# Returns to Physician Human Capital: Analyzing Patients Randomized to Physician Teams

**Joseph J. Doyle, Jr.**
**MIT & NBER**

**Steven M. Ewer, M.D.**
**University of Wisconsin—Madison**

**Todd H. Wagner**
**VA Palo Alto and Stanford**

**November 2008**

Abstract

Patient sorting can confound estimates of the returns to physician human capital. This paper compares nearly 30,000 patients who were randomly assigned to clinical teams from one of two academic institutions. One institution is among the top medical schools in the U.S., while the other institution is ranked lower in the quality distribution. Patients treated by the two teams have identical observable characteristics and have access to a single set of facilities and ancillary staff. Those treated by physicians from the higher-ranked institution have 10-25% shorter and less expensive stays than patients assigned to the lower-ranked institution. Health outcomes are not related to the physician team assignment, and the estimates are precise. Procedure differences across the teams are consistent with the ability of physicians in the lower-ranked institution to substitute time and diagnostic tests for the faster judgments of physicians from the top-ranked institution. The results suggest that more stringent minimum quality standards may lower health care costs.

*JEL Classifications:* I12; J24

The authors would like to thank Ann Bartel, Christine Durrance, Sherry Glied, Michael Greenstone, Larry Katz, Steve Levitt, Lars Lefgren, Abigail Ochberg, Joe Price, Roberto Rigobon, Jon Skinner, Doug Staiger, Tom Stoker, Tavneet Suri, and Jack Wennberg for helpful comments and insights. We also greatly benefited from the efficient programming of Andrew Siroka. All remaining errors are our own.

## 1. Introduction

Access to high-quality health care is a major social and economic issue in the U.S. Over $2 trillion is spent each year in the healthcare sector, and high-spending areas incur costs that are 50% higher than low-spending ones (Fisher et al., 2003). These differences are often ascribed to divergent preferences and training among physicians (Phelps and Mooney, 1993; Eisenberg, 2002), and there are equity concerns that health disparities may result from differences in access to high-quality care (Institute of Medicine, 2002; Chandra and Skinner, 2003; Almond, Chay, and Greenstone, 2008). State licensing requirements and medical school accreditation standards may further limit access to health care by raising prices.[1] A classic study by Friedman and Kuznets (1945), for example, attributed relatively high salaries among physicians, relative to dentists, to more stringent licensing requirements. They acknowledged that any effects on quality of care were beyond the scope of their investigation.

To better understand the effects of such minimum quality standards, it is necessary to estimate returns to physician human capital. There are two main limitations to estimating such returns: First, the environments where physicians operate may differ, including differences in complementary physical capital and human capital of the support staff. Second, high-risk patients may be referred to or self-select the "best" physicians (referral bias), and as a result the highest-quality physicians can have the highest mortality rates (Glance et al., 2008).[3] Indeed, public report cards that rank providers

[1] Kleiner (2000) provides a review.
[3] This non-random assignment of patients also plagues comparisons across hospitals. Geweke, Gowrisankaran, and Town (2003) show that patients with the worst unobservable severity go to high quality hospitals.

based on adjusted mortality rates have been controversial due to the concerns that patients differ in unobservable ways, and that the reports create incentives for providers to avoid high-risk cases (Marshall et al., 2000; Dranove et al., 2003).

This paper estimates returns to physician human capital by comparing treatment and health outcomes for patients treated by physicians that differ markedly in their levels of human capital. The main innovation considers a unique natural experiment in a large, urban Department of Veterans Affairs (VA) hospital, where nearly 30,000 patients (and over 70,000 admissions) were randomized to teams comprised of clinicians from one of two academic institutions. These clinical teams offer compelling variation in the human capital of the physicians: one institution is among the top medical schools in the U.S; the other is ranked lower in the quality distribution.

The two teams are composed of medical students, residents and attending physicians, but the residents serve the role of primary physician and thus their actions are most likely to contribute the bulk of any differences. Among residency programs, variations in delivery of health care can be explained both by differences in the quality of physicians accepted into the programs and in the quality of clinical training they receive during residency (Weiss, 1995; Semeijn et al., 2005). While it is not possible to separate the two, program curriculum, teaching philosophy and approach to clinical care are generally similar between the two institutions, and it is likely that differences in initial human capital levels account for a significant portion of any observed differences in health care delivery.

The empirical strategy employed in this paper offers two main advantages over previous research. First, the patient characteristics are identical across the two academic

institutions due to the randomization. Second, the teams have the same firm-specific human capital, with access to the same facilities, the same nursing staff, and the same specialists for consultations. The only difference is the physician team assigned to the patients. This allows a comparison of treatment decisions and health outcomes, controlling for patient characteristics and complementary physical and human capital.

We find that patients assigned to the higher-ranked program have 10% lower costs compared to the lower-ranked program, and up to 25% lower costs for more complicated conditions. The differences largely stem from diagnostic-testing rates. We find that the duration before the first test is longer for the lower-ranked institution, and that these physicians tend to order more tests once the first has been ordered. Meanwhile, hospital readmissions and mortality are unrelated to the physician-team assignment, and the estimates are precise. These results are consistent with the hypothesis that physicians in the lower-ranked institution successfully substitute time and diagnostic tests for the quicker judgment of the physicians in the higher-ranked institution. The evidence suggests that lowering minimum quality standards in medicine are less likely to reduce quality in terms of mortality or readmissions. However, if a greater proportion of patients are treated by physicians trained at higher-ranking institutions, then this could lower health care costs. [4]

A main caveat is that the results apply directly to one hospital, albeit with compelling variation in the physician characteristics. The parent hospital of the higher-

[4] The cost savings may be passed on to physicians in the form of higher wages in addition to the higher wage expected due to the restriction of supply. Unfortunately, detailed data linking wages to the quality of medical education do not exist.

[6] In the case of heterogeneous treatment effects, the patients are likely referred based on the expected gain of the assignment: a correlated random coefficient model that can inflate returns to physician human capital (Bjorklund and Moffitt, 1987).

ranked institution is similar in treatment intensity to other top teaching hospitals, however. This suggests that practice patterns at the top-ranked institution are similar to other highly ranked institutions as well.

The paper is organized as follows: section 2 describes the empirical framework and defines the main parameters of interest; section 3 provides background information on the physician teams and patient assignment, as well as a review of the previous literature; section 4 describes the data; section 5 reports the results; and section 6 concludes.

## 2. Empirical Framework

Consider a health production function that relates mortality, M, to health care inputs and a patient-level severity measure, $\theta$:

(1) $M = F(H, K; \theta)$

where $H$ represents human capital of the hospital staff, and $K$ represents physical capital.

The main parameter of interest here is the effect of physician human capital, $H$, on patient outcomes. In our empirical application, there are two teams that differ markedly in the screening of physicians that compose each team, including different residents and attending physicians. Let $P$ be an indicator that the patient was assigned to physicians in the lower-ranked program, $T$ be a measure of treatment, and $X$ represent observable characteristics of the patients. The main parameters of interest can then be written as:

$(2a)\ E(T \mid P = 1, X) - E(T \mid P = 0, X)$

$(3a)\ E(M \mid P = 1, X) - E(M \mid P = 0, X)$

This gives rise to empirical models of the form:

$$(2b)\ T_i = \alpha_0 + \alpha_1 P_i + \alpha_2 X_i + \varepsilon_i$$

$$(3b)\ M_i = \beta_0 + \beta_1 P_i + \beta_2 X_i + \upsilon_i$$

where $\varepsilon$ and $\upsilon$ are error terms.

A common problem when estimating $\alpha_1$ or $\beta_1$ is that patients are not randomly assigned to physicians. Rather, patients choose or are referred to physicians. A patient's primary physician, who knows more about the illness severity than can be captured in typical data sets, may refer the "toughest" cases to the "best" physicians. This tends to bias against finding survival improvements for physicians with higher levels of human capital.[6] Comparisons across hospitals have the additional confounding factors of differences in technology and support staff, which may have a large impact on patient survival independent of the physician characteristics (Unruh, 2003; Evans and Kim, forthcoming; Bartel, Phibbs, and Stone, 2008).

The main innovation in this paper is the study of a large number of patients who were randomly assigned to physician teams within the same facility. This should satisfy the identification assumptions that the physician team is mean independent of the error terms: $E(P\varepsilon) = E(P\upsilon) = 0$.

In terms of the standard errors, as in other randomized trials the individual error terms are assumed to be independently and identically distributed. The estimates reported are robust to heteroskedasticity and clustered at the patient level to account for dependence across observations for the same patients treated over time (similar results

are found when we restrict the analysis to each patient's initial episode, as described below). These errors are conservative compared to alternatives considered.[7]

## 3. Background

### A. Previous Literature

Much of the previous work on physician human capital finds that previous test scores, such as undergraduate grade point average or Medical College Admissions Test (MCAT) scores, are positively correlated with later test scores (Case and Swanson, 1993; Glaser et al., 1992; Hojat et al., 1997; Silver and Hodgson, 1997). It is less clear whether physicians with higher scores provide higher quality care. Ferguson et al. (2002) review the literature on predictors of medical school success, and note that little has been done on post-medical school performance. There is some evidence on outcome differences by board-certification status, but it is mixed.[8]

A measure of physician quality directly related to the current study comes from surveys of other physicians in the same market. Hartz et al. (1999) show that surgeons are more likely to be regarded as a "best doctor" in these community surveys if they

[7] One caveat is that the observations may be correlated within teams that vary over time, although we do not observe team composition. We found that clustering at the month-year level—times when the attending physicians are likely to change—resulted in similar, and often smaller, standard errors. Similarly, when the estimates were jointly estimated using a seemingly unrelated regression, estimated standard errors were again similar and often smaller. Last, we considered correlation within each of the two groups. The outcomes considered here, however, have an intra-class correlation of close to zero (e.g. our cost measures have an intra-class correlation of less than 0.005). As in other randomized trials, these intra-class correlation coefficients imply that correcting the standard errors by clustering at the group level is unnecessary in this context (Moulton, 1986; Angrist & Pischke, 2008).

[8] Certification has been found to be associated with reductions in mortality following heart attacks (Kelly and Hellinger, 1987; Norcini et al., 2000), while other work has found differences in the use of appropriate medications but little difference in mortality (Chen et al., 2006). Licensure examination scores have been found to be related to preventive care and more appropriate prescription medicines (Tamblyn et al., 1998; Tamblyn et al., 2002).

trained at a prestigious residency or fellowship program. They note that treatment by physicians trained at prestigious programs is not related to mortality, however.

Small-area variation in treatment has received considerable attention, with some evidence that physician quality measures vary across patient groups and may contribute to health disparities (see extensive reviews by van Ryn, (2002) and Bach et al. (2004)). In particular, access to high-quality specialists varies across racial groups, and desegregation has been found to significantly improve health outcomes for African American patients (Mukamel et al., 2000; Chandra and Skinner, 2003; Almond, Chay, and Greenstone, in press). Another reason for the large literature on small-area variation in treatment is that physicians are important cost drivers across areas. Physician characteristics have been found to explain up to 50% of the variation in expenditures, on par with case-mix variables (Pauly, 1978; Burns and Wholey, 1991; Burns, Chilingerian, and Wholey, 1994; Grytten and Sorensen, 2003).[9]

There is a related literature that considers the impact of report cards—publicly provided information about physician mortality rates, adjusted for case mix (for reviews, see Marshall, et al. (2000), Hofer et al., (1999), and discussions between Hannan and Chassin (2005) and Werner and Asch (2005)). Newhouse (1996) and Cutler et al. (2004) note that such report cards suffer from patient selection problems in ways that can confound estimates of the returns to physician human capital in general. For example, Dranove (2003) found limited access to surgery for high-risk patients following the

[9] Not all studies find significant effects of physicians on costs, however. Hayward et al. (1994) find that residents and attending physicians in one hospital do not explain much of the variation in length of stay (on the order of 1-2%).

introduction of report cards: fewer surgeries, more conducted at teaching hospitals, and large increases in adverse health outcomes in the short run.[10]

The empirical strategy in the literature to deal with these selection issues is a selection on observables approach—controlling for illness severity with indicators of comorbidities and patient characteristics such as age. Nevertheless, unobserved (to the researcher) differences in severity may contaminate comparisons. One study that is most similar to ours is an early study by Gillespie et al. (1989) that considered 119 patients randomized to two medical school programs in 1984 and 1985. They found little difference in diagnostic testing between the two programs. The analysis excluded patients who received no diagnostic testing, however, which may lead to sample selection bias. The current study will consider nearly 30,000 patients over 13 years. This includes over 72,000 patient encounters to provide a more comprehensive comparison, greater statistical power to detect differences, and a time frame that allows a comparison of long-term outcomes such as 5-year mortality.

**B. Training at the VA**

Physician training programs offer a way to accumulate human capital largely through learning by doing, and such training can have an effect on patient outcomes (Huckman and Barro, 2005).[11] One of the most common training grounds for physicians is the VA medical system.

The VA operates the largest integrated health care system in the US, with 155 medical centers and over 850 community-based outpatient clinics. Veterans can receive a range of services from general medical care to specialized services. In 2007, VA

[10] See also Schneider and Epstein (1996) and Omoigui (1996).
[11] See Marder and Hough (1983) for an early discussion on supply and demand for such opportunities.

treated over 5 million unique patients, and some health care reform experts use the VA as a model (Ibrahim, 2007]. The VA is organized around 21 regions, known as VISNs (Veterans Integrated Service Networks). Operating funds are distributed from Washington DC to each VISN, which then distributes the money to its hospitals, community clinics and outreach facilities. The financing system is based on a capitated risk-adjustment model.

Graduate medical education is part of the VA's statutory mission, and VA medical centers are located near academic medical centers to enhance training. 107 of the 126 medical schools in the U.S are affiliated with a VA medical center. The primary physicians for patients at VA hospitals are thus residents, particularly from internal medicine and general surgery training programs. Residents rotate through the VA system and treat many low income and disabled veterans—patients who provide valuable variation across a wide range of diseases. Each year, 31,000 residents (30% of all residents in the U.S.) and 17,000 medical students train in VA facilities (Chang, 2005; VHA, 2005).

This study considers a VA hospital in a large urban area that has affiliations with two medical schools.[12] This VA hospital is a full-service teaching hospital that provides over 3,500 surgical procedures each year. It has an intensive care unit and what are considered excellent laboratory facilities, including the ability to conduct magnetic resonance imaging and angiography. In addition to the main hospital, there are some smaller satellite hospitals elsewhere in the city that handle mental health, substance use treatment and long term care.

[12] We have chosen to keep the name of the VA hospital confidential out of respect for the patients and medical schools.

**C. The Residency Programs**

The variation in the medical and surgical residency training programs between the two institutions that serve this VA hospital is compelling: one is regarded as a top program in the U.S., whereas the other is ranked lower in the quality distribution. In the remainder of the paper, the higher-ranked institution will be referred to as Program A, and the lower-ranked institution will be referred to as Program B.

To establish the difference in credentials, Table 1 reports some summary characteristics of the two programs. First, the residency programs are affiliated with two different medical schools where the attending physicians that supervise and train the residents are faculty members. These medical schools differ in their rankings. Some years, the school affiliated with Program A is the top school in the nation when ranked by the incoming students' MCAT scores, and it is always near the top. In comparison, the lower-ranked program that serves this VA hospital is near the median of medical schools.

Another commonly used measure to compare medical schools is funding from the National Institutes of Health (NIH). This ranking identifies the major research-oriented medical schools, again with some of the most prestigious schools near the top. The medical school associated with Program A is again among the top schools in the U.S., whereas the lower-ranked program has an NIH funding level that is generally less than three out of every four medical schools.

Second, each training program is affiliated with another teaching hospital in the same city, in addition to the VA hospital. Program A's "parent hospital" is ranked among the top 10 hospitals in the country according the U.S. News and World Report

Honor Roll rankings of hospitals. Out of 15 specialties ranked by U.S. News, Program A's hospital is among the top 10 hospitals in the country for nearly half of them, and among the top 20 in nearly all of them (U.S. News & World Report, 2007). Meanwhile, Program B's parent hospital is not a member of this Honor Roll overall or ranked among the top hospitals in terms of subspecialties. The treatment intensity across the two parent hospitals is similar to one another, however, as described below.

Third, the residents themselves can be compared. Approximately 30% of residents who were trained in Program A received their M.D. from a medical school in the top 10 of the U.S. News and World Report rankings in 2004, compared to 3% of those trained in Program B. For top-25 medical schools, approximately half of Program A's residents graduated from such a school, compared to less than 10% for Program B. Similar differences are seen when the residents' medical schools are ranked by NIH funding levels. In addition, twice as many of Program B's physicians earned their medical degree from a medical school outside of the U.S.

At the end of the residency program students will often take board-certification exams, and the major Boards publish the pass rate for each residency program among those who were taking the exam for the first time. The two most relevant exams are given by the American Board of Internal Medicine and the American Board of Surgery. Table 1 shows that the pass rate for Internal Medicine is close to 100% for the residents in Program A compared to a pass rate of approximately 85% for Program B (a rate that is in the bottom quartile of the 391 programs listed).[13] The pass rate for General Surgery is

[13] American Board of Internal Medicine. Figures for 2005-2007. http://www.abim.org/pdf/pass-rates/residency-program-pass-rates.pdf

lower, 85% for Program A and 60% for Program B. These scores place Program A in the top quartile, and Program B in the bottom quartile, of residency programs in the U.S.[14]

In sum, it appears that the physicians in Program A perform substantially better on exams, and the affiliated medical schools differ markedly in prestige. These differences are stable over time, as a survey in the early 1970s asking medical school faculty to rank programs included Program A in its top 10, whereas Program B was ranked near the median of the rankings (Cole and Lipton, 1977).

**D. The Clinical Teams**

Discussions with physicians familiar with the programs revealed the similarities and differences across the teams. The clinical and teaching teams at this VA Medical Center conduct independent rounds each day during which they discuss their patients. The timing of these rounds does not differ systematically between the two institutions. This parallel structure allows a comparison of the two groups' treatment decisions and patient outcomes.[15] The patients assigned to each team are interspersed throughout the floors and share a common pool of nursing and ancillary staff. The two teams have access to the same specialists for consultations. There is a single set of clinical laboratories and imaging facilities for use by both teams, and conversations with physicians familiar with the operations at this hospital lead us to believe that neither institution receives favorable treatment from these ancillary service providers. We have also found that the overall philosophies of care do not differ substantially across the two

[14] American Board of Surgery, 5-year pass rate from 2002-2007. http://home.absurgery.org/default.jsp?prog_passreport

[15] Other VA Medical Centers that are served by multiple residency training programs generally allow the teams to mix, with rounds attended by all of the residents.

programs, and the amount of resident oversight at the VA is thought to be similar across the two programs.[16] This is described in more detail below.

Members of the clinical team include attending physicians, interns, senior residents and medical students, all of whom are affiliated with the parent teaching hospital. The intern, also known as a first-year resident, is the primary physician assigned to the patient, and this role includes evaluating patients, prescribing medicines, ordering diagnostic studies, performing bedside procedures, interacting with nursing staff and consultants, and writing the notes that make up the bulk of the medical record. The senior resident directly supervises the work of the intern, leads the team on daily rounds during which clinical care and teaching are afforded, and serves as a backup for the intern. The attending physician serves as the official provider of medial care and oversees the work of all other members of the team. This person typically does not attend the daily rounds of the team, but rather sees patients separately and discusses cases with the senior resident, confirming the clinical decision making of the team. Separate teaching rounds are provided for the team. The medical students, not yet physicians, are not allowed to write orders or officially contribute to the medical record. They work alongside residents to evaluate patients, and any contribution to decision making must go through the residents. This distribution of work is representative for teams in both Program A and Program B.

The size of the two physician teams is similar, consistent with the equal assignment of patients to the two teams. There are minor differences in the structure of the medicine teams, however. At a given time, Program A has four medicine teams, each

[16] Historically, VA hospitals were thought to provide less attending supervision than other teaching hospitals. In the 1990s, this was addressed and has continued to increase. For example, in 2004 the VA required an attending to be present for all major elective surgeries (Chang, 2005).

consisting of one attending physician, one senior resident and one intern. Program B likewise has four medicine teams composed of one attending and one senior resident, but these teams include two interns. This arrangement has remained consistent throughout the study period.[17] One explanation for faster treatment among the smaller teams in the higher-ranked program could be lower coordination costs, but teams do not coordinate care across the interns. In practice, the implication of this difference in team size is that Program B has an advantage in total residents (12 vs. 8). This favors Program B in terms of speed of treatment: a difference that would work against our main findings. Still, both teams have the same number of attending and senior residents. [STEVE: please check/edit above xxx]

**E. Patient Assignment**

To ensure an equitable distribution of cases and overall workload, the patients are randomly assigned to each institution: patients with social security numbers ending in an odd number are assigned to Program A and those with even social security numbers are assigned to Program B. This randomization method ensures that there is no crossover-if a patient is readmitted, the patient is assigned to the same physician group.

There are three exceptions to the randomization. First, the randomization only occurs at the main teaching facility, not at satellite facilities. Second, neurology patients are not randomized; rather all of the patients are assigned to one team. Third, the medical intensive care unit is headed by a single attending physician that oversees patients assigned to both teams. We will consider these groups of patients in specification checks below.

[17] Recently, Program B switched to a 3-team system described in an earlier version of this paper, but the change is outside of our sample period.

## 4. Data Description

We used the VA Patient Treatment Files (PTF) to identify inpatient encounters from 1993-2006. We restrict the main analysis to patients admitted to the main hospital facility, and patients who did not have a major diagnostic category of "nervous system"—these cases are less likely to enter the randomization. This results in an analysis data set of over 72,000 inpatient stays and nearly 30,000 patients. The main results include the information in all of the episodes and the standard errors are clustered by patient to take into account dependence within these observations as described above. Results will be shown for a sample restricted to patients' first episodes in the database as well.

The PTF includes the patient's age at admission, race, sex, marital status, and ZIP code of residence. Of these variables, the definition of race changed over time, as did its collection method (from admission-clerk assignment to self-report). This suggests that some caution is warranted with regard to this control. To corroborate the patient characteristics and control for neighborhood effects, data from the 2000 Census of Population were matched to the data to characterize the patient ZIP code, including the median household income, population density, and education, race, and age composition. Time and date of admission are also available, and the models include day-of-week, month, and year indicators, as well as indicators for 6-hour time-of-day blocks.

The PTF data also include ICD-9 diagnosis and procedure codes. This allows us to compare treatment across primary diagnoses, and 9 secondary diagnoses will be used to characterize the co-morbidities of the patient. It is possible that Programs A and B code diagnoses differently. This is testable in our data, as the sample sizes within

diagnoses can be compared across the 2 programs. These diagnosis codes are recorded for the benefit of patient histories and ongoing care rather than for billing purposes and, therefore, should not be affected by financial incentives to code patients into more profitable diagnoses (Dafny, 2005). Records can be coded by physicians or support staff, which would handle coding for both Programs A and B.

The VA PTF uses a scrambled social security number as the patient identifier. We linked this identifier to the last digit of the patient's true social security number to compare patients assigned to the different teams. The PTF does not have physician or resident identifiers to verify that all even numbered patients were indeed assigned to Program B, for example. After conversations with physicians familiar with the system, we do not expect patients with even-numbered social security numbers to be assigned to Program A apart from the exceptions listed in the background section.

There are four main measures of treatment provided. The patient's length of stay in the hospital is observed for all years in our dataset. Longer stays represent greater time for supervision and additional care. The VA strove to decrease length of stays in the mid-1990's by decentralizing power to geographic regions, changing ambulatory care benefits and creating incentives that reward medical center directors for shorter lengths of stay (Ashton et al., 2003). These policy changes would have been uniformly applicable to both Programs A and B, although we can test for differences in the response to these initiatives.

The second summary measure is the accounting cost of each stay. These data were not always included in the PTF, as the VA system provides care free of charge to veterans who have passed a means test or who have a service connected disability. The

VA has cost data after 1998 from the Decision Support System (DSS) and the Health Economics Resource Center databases. The DSS uses step-down accounting methods. Although the data are available for 1998 onward, we use DSS data from 2000-2006 when concerns about data completeness and accuracy were largely addressed.

The third summary measure is the Health Economics Resource Center Average Cost Data. These data are available from 1998 onwards, and uses non-VA (largely Medicare) relative value weights to estimate expenditures for VA care (Phibbs et al., 2003). One limitation of these estimated expenditures is that they are geared toward assigning average costs for patients with similar diagnoses and procedures, and are, therefore, less precise than DSS and can miss outlier costs (Wagner et al., 2003). Costs were standardized to 2006 dollars using the general urban consumer price index from the Bureau of Labor Statistics.

The fourth summary measure is the number and timing of procedures, based on ICD-9 procedure codes and dates. Physicians' use of diagnostic tests in particular can shed light on practice differences between Programs A and B.

There are two health outcomes that we consider. First, readmissions to the VA hospital within 30 days or 1 year of the date of admission are identified. A limitation of these readmissions is that they do not include admissions to non-VA hospitals. To the extent that lower quality care drives patients from the VA system and into a non-VA facility, then lower readmission rates could signal lower quality care. Still, many veterans depend on the free care provided by the VA, and we will generally regard readmissions as a negative outcome for patients. Another limitation is that any differences in initial length of stay will change the time at risk for a 30-day readmission,

for example. When the measure was 30-days from discharge (as opposed to days from admission), nearly identical results were found, however. Two related readmission measures consider the costs of these readmissions, and readmissions with the same major diagnosis as the initial episode.

The second outcome is more straightforward: mortality. The main results will focus on 30-day, 1-year, and 5-year mortality, and these measures were calculated for patients whose measures are not right censored. For example, 5-year morality was calculated for patients admitted to the VA hospital at least 5 years from the end of the sample period. These measures are taken from the VA vital status files and cover deaths occurring outside of the hospital as well as in-hospital mortality. These data have been shown to be highly accurate in terms of sensitivity and specificity (Arnold, et al., 2006). Other measures of mortality, such as 10-hour mortality, will be considered as well.

To describe the data available and compare patients assigned to the two groups, Table 2 reports summary statistics. The two columns of means are for patients with odd or even social security numbers: patients assigned to Program A and Program B, respectively. We do not believe that patients are aware of the dichotomy of physician teams and the difference in the quality of the residency programs, but to the extent that patients know they will be assigned to one of the two programs, sample selection could be an issue. If selection were a factor, then the observable characteristics may differ across the two groups as well as the frequency of observations.

Table 2 shows that the means are nearly identical across the two teams. Out of the 31 means comparisons, only 1 has a statistically significant difference at the 5% level, as expected by chance. Even this difference does not appear to be meaningful: among

Program B's patients, 27.1% of individuals in the average patient's ZIP code had some college education compared to 27.2% among those assigned to Program A. The average ages are nearly identical (63.0 and 62.8). The most common age is between 55 and 64, with smaller fractions of patients over the age of 65 when Medicare provides access to non-VA hospitals.[18] Still, there are many older patients in the sample, and the fraction of patients that no longer visit the VA hospital after the age of 65 does not vary systematically across the two physician teams.

Nearly all of the patients are male, an artifact of the older, veteran population. 47% are white, 44% are married, and 43% have a Charlson severity score of 2—an aggregation of the secondary diagnoses that is strongly associated with mortality (Quan et al., 2005). Most patients are admitted to the hospital between 12 noon and 6pm (42%), the average patient's ZIP code has a median household income of $34,000 and 63% of its population is white. The number of observations is similar across the two groups, with Program B treating 50.3% of the patients (35,932 vs. 36,434).[19] It appears that the patients who enter the VA hospital are randomly assigned to the two programs and that differential selection into the VA is unlikely to drive differences in treatment or health outcomes.

## 5. Results

### A. Treatment Differences

[18] Demand for VA care appears inelastic with regard costs of visiting a VA hospital. Mooney, et al. (2000) find that patients over the age of 65 are more inelastic with respect to distance to the VA hospital compared to those under the age of 65, despite access to Medicare for the older group.

[19] With the large sample size, this difference is marginally significantly different from 0.5 (p-value = 0.06). When first episodes are considered, the fraction assigned to Program B is 0.5002 (p-value = 0.92).

A first look at how the two programs' treatment levels differ can be seen in Figures 1A-1C. In each figure, the vertical axis reports one of the three summary measures of treatment: length of stay, accounting cost, and estimated expenditures. These data are right skewed and each measure was transformed using the natural logarithm. The means of the three measures are 1.43 log days (or 4.2 days), 8.63 log costs (or $5600 in 2006 dollars), and 8.71 log estimated expenditures (or $6000). The horizontal axis in each figure is the last digit of the patient's social security number. The last digit of the social security number is randomly assigned, and differences in the measures should stem solely from the difference in physician team assignment. Further, we would expect similar measures for each odd (or even) digit if differences in the physician team assignment were responsible for any differences as opposed to sampling variation.

Figures 1A-1C show a sawtooth pattern, with length of stay and the two cost measures 10 log points higher for patients with an even-numbered social security number compared to patients with an odd-numbered social security number; patients treated by Program B have higher costs. This difference is seen for each digit, as the means are similar for all even (or odd) last digits.

To aggregate the data up to the program level and introduce controls in the spirit of estimating equation (2b), Table 3 reports results from Ordinary Least Squares regressions for the three cost measures. Similar results were found when the length of stay was estimated as a count variable using a negative binomial model. Each column represents a separate regression. The first model reported includes no controls and the 10-11 log point differences shown in Figure 1 have a standard error of close to 1 log

point.[20] Results were similar, although slightly smaller, when the estimates were re-transformed and heteroskedasticity was taken into account (Manning, 1998).[21]

The second model includes 3-digit primary-diagnosis fixed effects to estimate differences in treatment within disease classes. These diagnoses may be affected by the choices of the physician teams, although this does not appear to be the case as described below. The models reported in Table 3 show that the results are largely unchanged when the diagnosis fixed effects are incorporated, although the estimates are slightly larger for accounting costs (12 log points).

The last column for each dependent variable includes the controls in Table 2, as well as year, month, and day-of-week indicators. The results are nearly identical to the model without the additional controls. This is consistent with the randomization effectively balancing the observable characteristics across the two groups, as shown in Table 2.

To place these results in context, Appendix Table A1 provides estimates for selected covariates. 10 log points is akin to an increase in age category from 45-54 to 65-69. Treatment measures in these data level off once the patient is 55, which may reflect a selection out of the VA hospitals once veterans are eligible for Medicare. Treatment levels for patients with a Charlson severity score of 2 are 11-13 log points higher compared to patients with a score of 1—a difference in severity that leads to substantial health outcome differences as described below. Admissions during business hours also

[20] The different samples for the cost measures are due to the different time periods when they are available.
[21] For models with full controls, when interpreting the estimates in terms of percentages rather than log points, a smearing factor (the ratio of the average exponentiated residuals in the regressions for each group) is applied and the estimated difference in length of stay is 10%; the difference in accounting cost is 9% and the difference in estimated expenditure is 8%.

accrue higher costs. Meanwhile, there is little relationship with day of admission, and married patients have 7-9% lower treatment levels compared to single patients.

Much of the remainder of the paper considers how the different programs differ in terms of procedures and across different types of patients to explore the mechanisms that drive the difference in the summary treatment measures. Before the sources of the treatment differences are explored, the next section reports tests of differences in health outcomes.

**B. Health Outcomes**

Given the results in Figure 1, it is possible that Program A discharges patients prematurely, and they may have worse long-term health outcomes. It is also possible that Program A provides higher quality care in less time and at lower expense. Figure 2 reports estimates of mean outcomes by the last digit of the social security number, and no differences are found across the patients in terms of 30-day readmissions, as well as 1-year and 5-year mortality.

Again to introduce controls and place the results in context, Table 4 reports the results of OLS regressions of the readmission and mortality indicators on the program assignment and controls (equation 3b). Results are similar when probit and logit models were used instead, partly because the dependent variables are sufficiently far from zero: 13% and 43% readmission rates at the 30-day and 1-year intervals, respectively, as well as 30-day, 1-year, and 5-year mortality rates of 6.4%, 24% and 51%.

Table 4 shows that the program assignment is unrelated to readmissions and mortality, with coefficients that are not statistically nor economically significant. For example, Program B is associated with a 0.6% increase in 1-year readmissions, or 1.4%

of the mean. When 1-year readmissions with the same major diagnostic code as the previous major diagnosis are compared, Program B is associated with a 0.3% increase or 1.5% of the mean.

In terms of mortality, Program B is associated with a 0.1 percentage-point reduction in 30-day mortality (or 1.1% of the mean), a 0.7 percentage-point reduction in 1-year mortality (or 2.9% of the mean), and a 0.3 percentage-point reduction in 5-year mortality (or 0.6% of the mean). The results are fairly precise as well. For 1-year mortality the 95% confidence interval is [-0.0155, 0.0016], and 5-year mortality the confidence interval is [-0.0162, 0.0106]. These differences are small compared to a 5-year mortality rate of over 50%, and largely rule out survival benefits from assignment to Program A. Across the 6 measures, the lower limit on the 95% confidence intervals are less than 7% of their respective means, and the upper limits are less than 5% of their means.

To place these small differences in mortality in context, other covariates are associated with higher mortality, as shown in Appendix Table A1. Men have 18% higher mortality rates, a Charlson severity score of 2 is associated with a 50% higher mortality compared to a score of 1, and mortality is strongly associated with the age of the patient.

## C. Mechanisms

### C.1. Diagnosis Complexity

To compare the robustness of the results across diagnoses and investigate whether the differences arise in more complex cases, Table 5 reports results from models estimated separately across common diagnoses. First, the top 10 most frequent diagnoses

are compared.[22] Two rows are presented for each diagnosis: estimates from a model for log length of stay—the resource measure that is available for the full time period, and 1-year mortality. Similar results were found for the other measures as well. The means of the dependent variables are listed, and they vary widely across the diagnoses.

The results show that for some serious conditions with high 1-year mortality rates, such as heart failure, chronic obstructive pulmonary disease (COPD), and pneumonia, treatment differences are between 20 and 25 log points. Smaller differences in treatment are found for less serious conditions such as chronic ischemic heart disease, with a difference closer to 10%. Acute myocardial infarction (AMI) has a 25% 1-year mortality rate, and a difference in log length of stay of 9 points.

To summarize all of the diagnoses, the 3-digit primary diagnosis codes were divided into quartiles based on their mortality rates.[23] No difference in treatment is found for the lowest quartile. This is a group with a 4% mortality rate and the treatment may be more standardized for less serious conditions. 11 and 12 log-point differences in length of stay are found for the 2nd and 3rd quartiles, and the most seriously ill patients have a 14 log-point difference in length of stay when the two Programs are compared. These cases are likely more complicated, as they have higher costs in addition to the higher mortality rates.

In terms of outcomes, the estimates are less precisely estimated within particular diagnoses given the smaller sample sizes, but the point estimates are unstable in sign and

[22] The top 10 diagnoses were determined by calculating the frequency of patients in 3-digit ICD-9 diagnosis codes, as well as more general definitions of gastrointestinal bleeding (Volpp et al., 2007) and Chronic Obstructive Pulmonary Disease.

[23] The mortality-rate quartiles could be affected by differences in the programs' diagnoses and their effectiveness, but when the conditions are scanned, they are similar to severity rankings when an independent dataset, the Nationwide Inpatient Sample, is used to characterize diagnoses by their mortality rates.

generally small in magnitude. The largest differences are found for AMI and cardiac dysrhythmias, with Program B associated with mortality rates that are 12-18% lower than the sample mean. These differences are not statistically significant, however, and no difference in 30-day readmissions is found for these diagnoses. In addition, no difference in 5-year mortality is found for AMI patients.[24] Program A is associated with lower mortality for pneumonia patients (5% lower compared to the sample mean); again the difference is not statistically significant. Across all of the other diagnosis categories, the hypothesis that Program A is associated with lower mortality is not found.

Table 5 also reports the fraction of patients treated by Program B for each diagnosis, along with a p-value from a test that the fraction of patients seen within a diagnosis equals 0.5. This tests whether the programs differ when recording the primary diagnosis. While some of the diagnoses show differences that are statistically significantly different from 0.5, all of the proportions are close to 0.5. In addition, the rates do not vary systematically with the mortality quartiles. It appears that the teams have similar primary diagnoses.

**C.2. Differences in Types of Care**

The summary measures of treatment can be disaggregated to better understand the types of care that differ across the two sets of physicians. Table 6 reports the results of 9 such models. The first is a simple count of the number of procedures, which averages 1.7. Patients assigned to Program B are found to receive 0.25 additional procedures on average. In terms of the types of procedures, column (2) shows that there is little

[24] For 30-day readmissions, the coefficient for the cardiac dysrhythmia sample is -0.006 compared to a mean of 13% and the coefficient for the AMI sample is -0.01 compared to a mean of 16%. The coefficient for 5-year mortality is -0.06 compared to a mean of 52% for cardiac dysrhythmias and -0.006 compared to a mean 49% for AMI.

difference in the number of surgeries. Much of the overall difference stems from differences in diagnostic procedures, and these differences will be explored further below.

The next six columns use the accounting cost segments, which sum to the total accounting cost measure described above. Levels (instead of logs) are used to avoid dropping observations with zero costs in a particular segment. Surgery costs are found to be $123 lower for Program B on average, or 9% of the sample mean. In all of the other categories, Nursing, Radiology, Lab, Pharmacy, and "all other" costs, Program B is associated with similarly higher costs in comparison to the mean for each segment, ranging from 7% of the mean for nursing care to 13% of the mean for laboratory costs.

One explanation for the lower costs associated with Program A is that these physicians may rely more heavily on outpatient care as a substitute for inpatient care. Our data describes whether an outpatient referral is made, which happens in most cases when a patient was admitted to the hospital (79% of the time). Program B is associated with a 1 percentage-point lower outpatient referral rate, which suggests that such substitution does not drive the inpatient cost differences.

**C.3. Differences in Diagnostic Testing**

To further explore the differences in diagnostic procedures, Table 7 reports rates of diagnostic testing across the two programs. Columns (1) and (2) report the frequency with which each program orders particular tests. For example, patients assigned to physicians from Program B are more likely to undergo diagnostic tests compared to patients treated by Program A (73% vs. 68%). This difference is found among common diagnostic tests including x-rays and stress tests. Columns (3) and (4) report the number of tests conditional on ordering any tests. Even conditional on ordering some tests,

Program B is found to order 8% more than Program A (3.25 vs. 2.99). Within procedures, the frequency of tests is more likely to be similar—a cardiac stress test, for example, is only conducted once (on average) in both groups if it is conducted at all.

Two potential explanations for the greater number of tests among Program B physicians are that they are less efficient in their decision making compared to the higher-ranked program, or they may receive training that stresses the importance of tests. One way to distinguish these explanations is to consider the time to the first test. If Program B has a stronger preference for ordering tests, they may order more tests and order them more quickly as well. If Program B takes more time to decide what course of action to take, or relies more heavily on input from consultants, then the time to the first test would be longer. Table 7 shows that the latter explanation is more likely: Program B is 10% slower, on average, to order the first test conditional on ordering one (1.55 days vs. 1.41). To account for the time at risk for procedures and include all observations, Cox proportional hazard models estimates show that for individual procedures, Program B is approximately 8% slower to order a test Program A. These differences are seen for x-rays, angiography, and cardiac tests.

The differences in Panel A may mask differences within particular diagnoses. 4 common diagnoses were chosen that have fairly standard diagnostic tests. The differences are less likely to be statistically significant due to the smaller sample sizes, but large point estimates point to patterns, especially the longer duration to the first test.

Panel B reports results for congestive heart failure, a chronic condition that is a common source of hospital admission. Higher test rates are found for Program B (5% higher overall; 19% higher for stress tests). Program B orders 14% more tests

conditional on any (3.33 vs. 2.92). In terms of timing, they take 21% longer to order the first test (1.34 days vs. 1.10 days), 51% longer to order an angiography if one is ordered (7.26 days vs. 4.81), 32% longer to order a cardiac stress test, and 74% longer to order other cardiac tests (including echocardiograms). Hazard ratios that take into account patients that did not receive the test as well show somewhat smaller but still economically and statistically significant differences: hazard ratios of 0.75 and 0.77 for angiography and cardiac stress tests, for example.

Panel C reports the results for myocardial infarction. Nearly every heart attack patient receives some diagnostic test, often an angiography. No difference is found in the rate of angiography across the two programs, but Program B takes 10% longer to have one conducted. Program B is associated with 40% higher rates of cardiac stress tests (30% vs. 21%) and higher rates of "other cardiac tests including echocardiograms. Conditional on ordering the tests, they order 8% more and have a 7% longer duration to the first test, including 50% more time before tests such as an echocardiogram is taken (3 days vs. 2 days). The hazard ratios are closer to 0.90 for angiography and other cardiac tests.

Panel D reports the results for another common admission: chronic obstructive pulmonary disease. Overall, diagnostic-testing rates are similar across the programs, although Program B is 17% more likely to order a chest x-ray and 13% more likely to order any x-ray compared to Program A. The main difference within this diagnosis is the time to the first test: 59% longer for Program B on average (0.94 days vs. 0.59 days), and approximately 25% longer for an x-ray (hazard ratios of 0.91 and 0.82). Panel E reports similar results for gastrointestinal bleeding, with 6% higher test rates, 11% more tests

conditional on ordering any, and 27% longer duration before the first test (0.94 days vs. 0.74 days), with a hazard ratio for endoscopy of 0.85.

In summary, Program B orders more diagnostic tests, even conditional on ordering any tests. This is consistent with a group that is either more careful or a group that requires more time and information to understand the nature of the condition. The shorter duration before the first test is suggestive that Program A is faster at determining the nature of the health problem. Although we are not able to directly measure it, an increased reliance on subspecialty consultation by Program B could contribute to these findings, as well.

**D. Robustness & Specification Checks**

This section considers alternative explanations for the main results and uses the structure of the natural experiment to examine whether these explanations can account for the differences in treatment, as opposed to differences in physician human capital. Most of the results are presented in Table 8, which reports a number of specification and robustness checks. Each row represents a separate model with full controls.

**D.1. Placebo Tests**

The first set of results compares patients who were not subject to the randomization. Patients directly admitted to the Neurology service, such as stroke patients, are not randomized to the two teams. When the major diagnostic category of "nervous system" patients were considered—a group that is less likely to enter the randomization—a much smaller treatment difference is found (coefficient of 0.047), and the difference is not statistically significant. Second, when patients admitted to a satellite facility (where randomization does not take place) were considered, again there is no

difference in length of stay or 1-year mortality. These results are consistent with the idea that patients with odd or even social security numbers are similar to one another, including their propensity to visit the VA or to receive care at a satellite clinic.

The other area where the randomization has less of an effect is when a patient is admitted to the intensive care unit, which is overseen by a single attending from one of the programs at any given point in time. There were no differences for patients who were admitted to the ICU in terms of length of stay in the ICU or mortality in the ICU. For patients who were transferred out of the ICU to another hospital bed, their post-ICU length of stay was significantly different.

Further, when patients who did not use an intensive care unit were analyzed, the treatment differences were somewhat larger in magnitude, and no outcome differences were found. We also did not find a difference in the rate of transfer to the ICU across the two groups.

**D.2. Initial episodes**

Given little difference in readmissions, we chose to use the information from all of the patient encounters in the main results. Perhaps a cleaner measure of treatment and outcome differences can be found by looking at the patient's first episode of care. Treatment differences are similar to the main results when the sample is restricted in this way. In terms of outcomes, the coefficient on assignment to Program B for the 30-day readmission model increases in magnitude to -0.010, and the result is statistically significant. As noted above, however, the readmission variable is somewhat problematic given the censoring of readmissions to non-VA facilities and the different lengths of time that the patients are at risk for readmission given the longer initial stay lengths for

Program B. When 1-year mortality is considered instead, the coefficient decreases in magnitude to -0.004, or 2.3% of the sample mean. A similar coefficient was found for 5-year mortality, or 0.8% of the mean.

**D.3. Heterogeneity Across Patients**

Part of the interest in estimating the returns to physician human capital is the concern that minority patients may lack access to top physicians. The natural experiment here allows us to compare the treatment and outcome differences for white vs. non-white patients, although the non-white category includes missing race (Sohn et al., 2006). Racial composition in the patent's ZIP code is associated with the race listed in the patient treatment file, however, which suggests that the race variable is informative.[25] Table 8 shows that the difference in treatment is larger for non-white patients (14 log point difference in length of stay compared to 8 log points for white patients). 1-year mortality is similar across whites and non-whites at 24%, and the Program assignment is unrelated to this outcome.

The main results include controls for age categories, and the results are similar when individual age indicators are used, as expected given the randomization. Further, the distribution of ages suggests that once individuals turn 65, some may opt for non-VA care due to the availability of Medicare. Similar results were found when the models were separately estimated for individuals under the age of 65 and over the age of 65. Results were also similar when the analysis was conducted from 1993-2000 and 2000-2006, with somewhat larger treatment differences in the latter period.

[25] We divided the sample into quartiles based upon the fraction white in the patient's ZIP code. Patients in the bottom quartile are recorded as white 9.5% of the time compared to 72% in the top quartile. When treatment and outcomes are compared, the bottom quartile shows the largest difference in log length of stay (16 log points), and a model without controls suggests that Program B is associated with mortality that is 2 percentage points lower compared to a mean of 25% in this quartile.

### D.4. Alternative Outcome Measures

The last set of rows in Table 8 report outcome results, and the results are robust. First, 30-day readmissions for the same major diagnostic category measures re-hospitalizations that are more directly related to the initial admission. The next set of outcomes considers the time period when the estimated expenditures are available, and the costs associated with the readmission are used as a measure of severity. Levels instead of logs are used to retain the information included from patients with no readmissions. Readmission costs are found to be only $20 higher for patients assigned to Program B compared to a mean of over $1650. Similarly, for 1-year readmission costs, Program B is associated with $240 higher costs on average, compared to a mean of nearly $5000.

Another mortality measure that perhaps has the most direct influence of the resident team is mortality in the hospital. This could be due to more aggressive surgical tendencies or lower quality care. The in-hospital mortality rate for all diagnoses is 4%, and Program B is associated with a 0.2% higher mortality rate, a difference that is not statistically significant.

One implication of the difference in the timing of diagnostic tests is that Program B substitutes time for quicker decision making. We find that this is not related to 30-day or 1-year mortality. An instance where delay in decision making may be crucial is mortality in the first few hours. Table 8 shows that 10-hour mortality (from the time of admission into the hospital) is similar across the two programs, however. Similar results are found for 5-hour mortality and 1-day mortality, as well as in-hospital mortality.

An explanation for the shorter stays associated with Program A could be that these physicians are more likely to transfer patients to another hospital, potentially to perform a surgery that is not conducted at the VA such as a coronary artery bypass. Table 8 shows that Program B is associated with a slightly lower transfer rate: 0.3% compared to a mean transfer rate of 4%. This difference cannot by itself explain the difference in length of stay.[26] Further, when (the small number of) transferred patients were dropped from the analysis, the results are essentially the same as the main results (see Appendix Table A2).

**D.5. June vs. July: Heterogeneity in Resident Experience**

One limitation of the analysis of residents is that the practice styles and outcomes may converge or diverge as the physicians gain experience later in their careers. Future analysis will use Medicare data to track these physicians into the future, where the adequacy of patient controls to mimic the randomization will be tested by estimating models of the effect of these physicians on treatment and outcomes shortly after the residency. In these data, we can compare patients in June versus July—the month when new residents begin training and the pool of residents has nearly one less year of experience. This two-month comparison also controls for seasonal differences in the types of conditions encountered. Given the smaller sample sizes, results should be taken with some caution, as the differences between June and July are not statistically significant. That said, we find that the magnitude of the treatment differences is smaller in June when the residents are more seasoned (7% difference). Patients assigned to Program B when the residents are relatively inexperienced in July have lengths-of-stay

[26] For this difference in transfer rate to explain the 10% difference in length of stay, those patients more likely to remain due to Program B assignment would have to stay for 139 days compared to a mean of 4.4.

that are an additional 5% longer (see Appendix Table A3). The outcome results are more mixed: readmissions for the inexperienced July residents in Program B are higher than those in June, although July patients assigned to Program A are found to have slightly lower readmission rates. Mortality differences are also small and not statistically significant.

**D.6. Differences When Workloads Differ**

Another test used the idea that at times there would happen to be a number of patients admitted with even or odd-numbered social security numbers. This provides a test of the effect of workload on outcomes, and the results can be compared across the programs. Each team sees approximately 50 patients per week on average. Busier times were generally associated with healthier patients in terms of lower mortality rates. We do not find a significant interaction between the number of patients at a given point in time and the physician team assigned for treatment or outcomes.

**D.7. Additional Robustness Tests**

Other tests were conducted that are not shown in Table 8. Results were similar when date fixed effects were used to compare patients within the same date of admission to control for differences that may vary over the course of the year with different rotations. Probit models also yielded similar results (see appendix Table A2). In addition, the data contain admission and discharge times, so hours in care can be examined. We find that Program B is associated with 10% more hours in care (an average of 14 hours compared to a mean of 140 hours). It appears that the main results are robust to alternative specifications and samples, consistent with the ability of the randomization to control for unobservable characteristics.

## E. Interpretation

### E.1. Competing Explanations

These two training institutions differ in their level of academic prestige, a finding that is consistently supported by several different metrics. It is not possible to completely separate the difference in the baseline characteristics of those who gain admission to the residency programs versus differences in quality of training once in the programs.

There are a number of explanations for the treatment differences and outcome similarities. One is that Program A is more efficient at determining the proper treatment, possibly relying less on consultants to determine the clinical course. This explanation is supported by the larger differences in treatment for more complicated diagnoses, the larger number of diagnostic tests ordered by Program B, and the longer duration before the first test for Program B.

Another potential explanation is that the training styles of the two groups may differ. This does not appear to drive the results, however. It is not the case that Program B trains in a "parent hospital" that stresses extra time in care or a greater number of tests. According to the Dartmouth Atlas performance reports for 2001-2005, the average hospital days per Medicare beneficiary during the last two years of life—a preferred measure of utilization that controls for the health of the patient and is not directly affected by price differences—is nearly identical for the two parent hospitals. They also have similar facility capacity in terms of total beds and ICU beds—measures that have been found to be associated with treatment intensity (Fisher et al., 1994). If anything, the Medicare reimbursements for procedures, imaging, and tests are higher for the parent

hospital of Program A, although it appears that Program A has higher prices rather than differences in quantity of care in general. These results suggest that differences in treatment philosophies are not driving the treatment differences. In addition, conversations with physicians familiar with the two programs reveal little difference in the treatment philosophies between the two programs.

A related explanation is that the attending physicians in Program B provide more oversight, which takes more time to administer. If a mechanical rule that all tests had to be approved by the attending led to the cost differences, we would expect differences in treatment even for less serious cases, but that was not found (Table 5). In some ways, additional supervision may capture important differences in the two programs if the physicians in the lower-ranked program require additional advice. Again, physicians familiar with the training at this VA do not believe that the level of attending supervision is significantly different across the two groups, although such differences cannot be entirely ruled out. In the end, patients were assigned to either a particularly prestigious physician team or a less prestigious one, and those teams are provide a bundle of characteristics including the attending physicians who are on faculty at medical schools that vary widely across various rankings, as well as the residents who treat the patients directly.

**E.2. Implications**

The American Medical Association is a textbook example of the use of accreditation standards to limit the supply of physicians. This provides a minimum quality standard and is thought to increase prices. While caution is warranted when extrapolating the current results, they suggest that a relaxation of these standards would

not have adverse effects on health outcomes. Further, a relaxation of the standard—and an increase in the supply of lower-skilled physicians—may not lower costs due to a countervailing effect of higher treatment intensity. One question that arises is whether the cost savings associated with physicians who trained at prestigious programs, if persistent, are passed on in the form of higher wages. Unfortunately, physician-income surveys do not include physicians' medical school or residency training program to test such a relationship.

Another issue is scalability. Given the lack of a difference in health outcomes, it appears that physicians in Program B successfully substitute time and diagnostic tests for skills associated with admission into and training received from the higher-ranked program. One possibility is that the physicians in the lower-ranked program have identical initial reactions as to the proper course of action but are less confident in their initial judgments. If this were the case, it would be possible for the lower-ranked physicians to achieve similar outcomes at substantial savings. To the extent that physicians need more time for additional testing and input from consultants to achieve the same results as the higher-ranked program, the decision-making ability of the physicians in the higher-ranked program would not be scalable.[28]

We investigated these possibilities in two ways. First, we considered the admission diagnosis versus the settled-upon principal diagnosis, but we did not find differences across the two programs in the admission diagnosis. This is consistent with

[28] In some ways the top-ranked program's physicians are "stars". Rosen (1981) discusses star physicians, where the potential to be a superstar is limited by the extent of the market—in this case the physician's time to see patients. This time constraint inhibits the scalability of the treatment provided by top physicians.

the notion that the overall diagnosis is not related to the underlying skill of the physician in the vast majority of cases. Second, if the physicians all had the same initial interpretation of the patient's condition, but the physicians in the lower-ranked program were taking time and ordering tests to corroborate the initial impression (while the physicians in the higher-ranked program were more confident in the initial interpretation), then the initial tests of the two groups of physicians should be similar and differences should only arise in subsequent tests. We find that the time to the initial test differs across the groups, however. This suggests that the greater efficiency of the higher-ranked program is less likely to be scalable.

**E.3. Limitations**

There are a number of limitations in the current study. Perhaps most important is that the randomization applies to two sets of residency programs. While the variation in the programs is compelling, there is a question of external validity. One reason to believe that there may be wider applicability is that Program A's parent hospital is fairly similar to other U.S. News and World Report's Honor Roll Hospitals according to the Dartmouth Atlas. In terms of average number of hospital days and the number of physician visits in the last two years of life between 2001 and 2005, the parent hospital is in the middle of the distribution of these hospitals. It appears that other top hospitals provide similar levels of treatment intensity as the higher-ranked program. As noted above, the parent hospital affiliated with Program B has similar treatment intensity measures as the parent hospital for Program A—both are higher than the national average, but not at the extremes like some Honor Roll hospitals.[29]

[29] We thank Jack Wennberg for this suggestion.

A second limitation is that to the extent that the results are driven by different residents, as opposed to different attending physicians, then the differences could fade (or increase) over time as physicians gain experience. The June vs. July comparisons described above suggest that treatment differences may converge somewhat, although the outcome differences were similar when the residents were relatively inexperienced.[30] While residents may differ from more experienced physicians, one study found their practice patterns to be similar: Detsky et al. (1986) examined a strike by residents in 1980 and found that the volume of tests performed did not change when the attendings provided the care instead.

Third, the results apply to a veteran population, and the results may not apply to a wider set of patients. Still, this population is particularly policy relevant given the concerns that differing access to high-quality physicians may lead to health disparities among low-income groups. Here, we have just such a group that has an equal chance of being treated by a top physician team or one ranked much lower. Further, medical schools join with VA medical centers partly because the patients present with a wide range of illnesses—an advantage here in that we can compare the results across these diagnoses as well.

Further, a usual limitation of randomized trials is that they do not incorporate the value of matching physicians to patients. Here, the lack of a health outcome difference suggests that such triage is less likely to be necessary. In addition, to the extent that the

[30] Future research will test whether the differences among these residents can be shown in their early years as attending physicians, using patient controls in a "selection on observables" strategy. To the extent that these controls can mimic the randomization used in the main analysis, tests of whether the differences in treatment converge or diverge will be conducted, as well as tests of whether differences in outcomes emerge over time.
.

cost savings would be greater with matching, the magnitude of the cost-savings we find associated with treatment by a highly-ranked physician team can be viewed as a lower bound.

## 6. Conclusions

Physicians play a major role in determining the cost of health care, and there are concerns that limitations on the supply of physicians and disparities in access to high-quality physicians and facilities can affect health outcomes. Comparisons of physicians are often confounded by differences in the patients they treat and the environments where they work. We study a unique natural experiment where nearly 30,000 patients were randomized to two physician teams in the same hospital. The two teams are affiliated with academic institutions that differ markedly in prestige. One has residency programs that are consistently ranked among the top programs in the country, whereas the other has training programs ranked lower in the quality distribution according to measures such as the pass rate for Board exams.

We find patients randomly assigned to the higher-ranked program incur substantially lower costs: 10% overall and up to 25% depending on the condition. This difference is driven largely by variation in diagnostic testing, where Program B orders more tests and takes longer to order them. No difference is found for health outcomes, however. The results suggest that a relaxation of accreditation standards—to the extent that new physicians are similar to those who trained at the lower-ranking institution studied here—would not adversely affect quality of care, but may adversely affect costs due to greater average treatment intensity among physicians trained at lower-ranking institutions.

These results do not appear to stem from differences in training styles or treatment philosophies across the two programs. Rather, the results are consistent with physicians in the lower-ranked program successfully substituting time and diagnostic tests for the faster treatment associated with the higher-ranked program.

**References**


Angrist, Joshua D. and Jorn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. New Jersey: Princeton University Press.

Almond, Douglas, Chay, Kenneth, and Michael Greenstone. (forthcoming) "Civil Rights, the War on Poverty, and Black-White Convergence in Infant Mortality in the Rural South and Mississippi" *American Economic Review*.

Arnold N., Sohn M., Maynard C., and D.M. Hynes. 2006. "VA-NDI Mortality Data Merge Project." *VIReC Technical Report 2*. Edward Hines, Jr. VA Hospital, Hines, IL: VA Information Resource Center.

Ashton CM, Souchek J, Petersen NJ, Menke TJ, Collins TC, Kizer KW, Wright SM, Wray NP. 2003. "Hospital Use and Survival among Veterans Affairs Beneficiaries." *New England Journal of Medicine*. 349(17): 1637-1646.

Bach, P.B., Phram, H.H., Schrag, D., Tate, R.C., and J.L. Hargraves. 2004. "Primary Care Physicians who Treat Blacks and Whites." *New England Journal of* Medicine. 351(6): 575-584.

Bjorklund, Anders and Robert Moffitt. "The Estimation of Wage Gains and Welfare Gins in Self-Selection Models." The Review of Economics and Statistics. 69(1). February 1987: 42-49.

Burns Lawton R., and Douglas R. Wholey. 1991. "The Effects of Patient, Hospital, and Physician Characteristics on Length of Stay and Mortality." *Medical Care*. 29(3): 251-271.

Burns, Lawton R., Chilingerian, Jon A. and Douglas R. Wholey. 1994. "The Effect of Physician Practice Organization on Efficient Utilization of Hospital Resources." *Health Services Research*. 29(5): 583-603.

Case SM, Swanson DB. (1993) Validity of the NBME Part I and Part II scores for selection of residents in orthopaedic surgery, dermatology, and preventive medicine. *Academic Medicine. 68*:S51–S56.

Chandra, Amitabh, and Jonathan Skinner. (2003) "Geography and Racial Health Disparities," NBER Working Paper No. 9513.

Chang, Barbara K. (2005). "Resident Supervision in VA Teaching Hospitals." *ACGME Bulletin*. September: 12-13.

Chen, Jersey, Rathore, Saif S., Wang, Yongfei, Radford, Martha J., and Harlan M. Krumholz. (2006) "Physician Board Certification and the Care and Outcomes of Elderly Patients with Acute Myocardial Infarction. *Journal of General Internal Medicine*. 21(3): 238-244.

Cole, Jonathan R. and James A. Lipton. 1977. "The Reputations of American Medical Schools." *Social Forces.* 53(3): 662-684.

Cutler, David M., Huckman, Robert S., and Mary Beth Landrum. 2004. "The Role of Information in Medical Markets: An Analysis of Publicly Reported Outcomes in Cardiac Surgery." *American Economic Review*. 94(2) Papers and Proceedings of the One Hundred Sixteenth Annual Meeting of the American Economic Association. May: 342-346.

Dafny, Leemore. 2005. "How Do Hospitals Respond to Price Changes." *American Economic Review*. December. 95(5): 1525-1547.

Detsky, Allan S., McLaughlin, John R., Abrams, Howard B., L'Abbe, Kristan, and Frank M. Markel. 1986. "Do Interns and Residents Order More Tests than Attending Staff? Results of a House Staff Strike." *Medical Care*. 24(6): 526-534.

Dranove, David, Kessler, Daniel, McClellan, Mark, and Mark Satterthwaite. 2003. "Is More Information Better? The Effects of "Report Cards" on Health Care Providers." *Journal of Political Economy*. 111(3): 555-588.

Eisenberg, John M. 2002. "Physician Utilization: The State of Research about Physicians' Practice Patterns." *Medical Care*. 40(11): 1016-1035.

Evans, William N., and Beom Soo Kim. Forthcoming. "Patient Outcomes When Hospitals Experience a Surge in Admissions." *Journal of Health Economics*.

Fisher, Elliott S., Wennberg, John E., Stukel, Therese A., Sharp, Sandra M. (1994) "Hospital Readmission Rates for Cohorts of Medicare Beneficiaries in Boston and New Haven" *New England Journal of Medicine.* 331: 989-995.

Fisher E, Wennberg D, Stukel T, Gottlieb D, Lucas F, Pinder E. (2003) "Implications of regional variations in Medicare spending. part 2: health outcomes and satisfaction with care." *Annals of Internal Medicine* 138(4): 288-298.

Ferguson, Eamonn, James, David, and Laura Madeley. 2002. *British Medical Journal* "Factors Associated with Success in Medical School: Systematic Review of the Literature." 324: 952-957.

Friedman, Milton and Simon Kuznets. 1945. *Income from Independent Professional Practice*. New York: National Bureau of Economic Research.

Geweke, J. Gowrisankaran, G, and R.J. Town. 2003. "Bayesian Inference for Hospital Quality in a Selection Model." *Econometrica* 71(4): 1215-1238.

Gillespie, Kathleen N., Romeis, James C., Virgo, Kathy S., Fletcher, James W., and Anne Elixhauser. (1989) "Practice Pattern Variation Between Two Medical Schools." *Medical Care* 27(5):537-542

Glance, Laurent G., Dick, Andrew, Mukamel, Dana B., Li, Yue, and Turner M. Osler. 2008. "Are High-Quality Cardiac Surgeons Less Likely to Operate on High-Risk Patients Compared to Low-Quality Surgeons? Evidence from New York State." *Health Services Research*. 43(1): 300-312.

Glaser K., Hojat M., Velkoski J.J., Blacllow R.S., and C.E. Goepp. (1992) "Science, verbal, or quantitative skills: which is the most important predictor of physician competence?" *Educational and Psychological Measurement* . 52:395-406

Gowrisankaran, Gautam and Robert J. Town (1999). "Estimating the Quality of Care in Hospitals Using Instrumental Variables," *Journal of Health Economics* 18: 747 – 67.

Grytten, Jostein and Rune Sorensen. 2003. "Practice Variation and Physician-Specific Effects." *Journal of Health Economics*. 22: 403-418.

Hannan, Edward L. and Mark R. Chassin. 2005. "Publicly Reporting Quality Information." 293(24): 2999-3000.

Hartz. Arthur J., Kuhn, Evelyn M., and Jose Pulido. 1999. "Prestige of Training Programs and Experience of Bypass Surgeons as Factors in Adjusted Patient Mortality Rates." *Medical Care*. 37(1): 93-103.

Hayward, Rodney A., Manning, Jr., Willard G., McMahon, Jr., Laurence F., and Annette M. Bernard. 1994. "Do Attending and Resident Physician Practice Styles

Account for Variations in Hospital Resource Use?" *Medical Care*. 32(8): 788-794.

Hofer, Timothy P. et al. 1999. "The Unreliability of Individual Physician 'Report Cards' for Assessing the Costs and Quality of Care of a Chronic Disease." *Journal of the American Medical Association*. 281: 2098-2105.

Hojat, Mohmmadrez, Gonnella, Joseph S., Erdmann, James B., and J. Jon Veloski. 1997. "The Fate of Medical Students with Different Levels of Knowledge: Are the Basic Medical Sciences Relevant to Physician Competence." *Advances in Health Sciences Education*. 1: 179-196.

Huckman, Robert and Jason Barro. 2005. "Cohort Turnover and Productivity: The July Phenomenon in Teaching Hospitals." *NBER Working Paper*. No. 11182.

Ibrahim SA. 2007. The Veterans Health Administration: a domestic model for a national health care system? *American Journal of Public Health*. December. 97(12): 2124-2126.

Institute of Medicine. 2002. "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care." Washington D.C.: National Academies Press.

Kelly, J.V. and F.J. Hellinger. 1987. "Heart Disease and Hospital Deaths: An Empirical Study." *Health Services Research*. 22(3) August: 369-95.

Kleiner, Morris M. 2000. "Occupational Licensing." *Journal of Economic Perspectives*. 14(4): 189-202.

Manning, Willard G. 1998. "The Logged Dependent Variable, Heteroscedasticity, and the Retransformation Problem." *Journal of Health Economics*. 17: 283-295.

Marshall, Martin N., Shekelle, Paul G., Leatherman, Sheila, and Robert H. Brook. 2000. "The Public Release of Performance Data: What Do We Expect to Gain? A Review of the Evidence." *Journal of the American Medical Association*. 283: 1866-1874.

Mooney, C., Zwanziger, J., Phibbs, C., and S. Schmitt. 2000. "Is Travel Distance A Barrier to Veterans' Use of VA Hospitals for Medical Surgical Care?" *Social Science and Medicine*. 50(12): 1743-1755.

Moulton, Brent. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics*. 32: 385-397.

Mukamel, Dana B., Murthy, Ananthram S., and David L. Weimer. (November 2000) "Racial Differences in Access to High-Quality Cardiac Surgeons." *American Journal of Public Health*. 90(11): 1774- 1777.

Newhouse, Joseph. 1996. "Reimbursing Health Plans and Health Providers: Efficiency in Production vs. Selection." Journal of Economic Literature. 34(3): 1236-1263.

Norcini JJ, Kimball HR, and Lipner RS. (2000) "Certification and specialization: do they matter in the outcome of acute myocardial infarction?" Academic Medicine. 75:1193–1198.

Omoigui, Nowamagbe A., Miller, Dave P., Brown, Kimberly J., Annan, Kingsley, Cosgrove, Delos, Bytle, Bruce, Loop Floyd, and Eric J. Topol. 1996. "Outmigration for Coronary Bypass Surgery in an Era of Public Dissemination of Clinical Outcomes." Circulation. 93(1): 27-33.

Pauly, Mark V. 1978. "Medical Staff Characteristics and Hospital Costs." Journal of Human Resources. 13(S): 77-111.

Phelps, Charles and Cathleen Mooney. 1993. "Variations in Medical Practice Use: Causes and Consequences." In *Competitive Approaches to Health Care Reform*, ed. Arnould, Richard, Rich, Robert, and William White. Washington DC: The Urban Institute Press.

Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, and WA Ghali. (2005) "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data." *Medical Care* 43(11):1073-1077.

Phibbs, Ciaran S., Bhandari, Aman, Yu, Wei, and Paul G. Barnett. 2003. "Estimating the Costs of VA Ambulatory Care." *Medical Care Research and Review*. 60(3): 54S-73S.

Rosen, Sherwin. 1981. "The Economics of Superstars." *American Economic Review*. 71(5): 845-858.

Schneider, Eric C. and Arnold M. Epstein. 1996. "Influence of Cardiac Surgery Performance Reports on Referral Practices and Access to Care." *New England Journal of Medicine*. 335(4): 251-256.

Semeijn, Judith, Van der Velden, Rolf, Heijke, Hans, Van der Vleuten, Cees, and Henny Boshuizen. 2005. "The Role of Education in Selection and Allocation in the Labour Market: An Empirical Study in the Medical Field." *Education Economics*. 13(4): 449-477.

Silver, B. and C.S. Hodgson. (1997) Evaluating GPAs and MCAT scores as predictors of NBME I and clerkship performances based on students' data from one undergraduate institution. 72(5): 394-396.

Sohn, M.W., Arnold, N., Maynard, C., and D.M. Hynes. 2006. "Accuracy and Completeness of Mortality Data in the Department of Veterans Affairs." *Population Health Metrics.* 4(2): available at: http://www.pophealthmetrics.com/content/4/1/2.

Tamblyn, Robyn, Abrahamowicz, Michael, Brailovsky, Carlos, Grand'Maison, Paul, Lescop, Joelle, Norcini, John, Girard, Nadyne, and Jeannie Haggerty. (1998) "Association Between Licensing Examination Scores and Resource Use and Quality of Care in Primary Care Practice" JAMA 280:989-996.

Tamblyn Robyn, Abrahamowicz Michael., Dauphinee D., Hanley J.A., Norcini John, Girard Nadyne, Grand'Maison Paul, and Carlos Brailovsky. (2002) "Association between Licensure Examination Scores and Practice in Primary Care." JAMA 288: 3019-3026.

Unruh, L. 2003. "Licensed Nurse Staffing and Adverse Events in Hospitals." Medical Care. 41: 142-152.

U.S. News and World Report. 2007. "Best Hospitals 2007." Accessed via web at: http://health.usnews.com/usnews/health/best-hospitals/honorroll.htm

Van Ryn, M. 2002. "Research on the Provider Contribution to Race/Ethnicity Disparities in Medical Care." Medical Care. 40(1): 140-151.

Volpp, Kevin G., Rosen, Amy K., Rosenbaum, Paul R., Romano, Patrick S., Even-Shoshan, Orit, Wang, Yanli, Bellini, Lisa, Behringer, Tiffany, and Jeffrey H. Silber. 2007. "Mortality Among Hospitalized Medicare Beneficiaries in the First 2 Years Following ACGME Resident Duty Hour Reform." *Journal of the American Medical Association*. 298:975-983.

VHA (Veterans Health Administration). 2005 "Report to the Secretary of Veterans Affairs". Accessed at http://www.va.gov/oaa/archive/FACA_Report_2005.pdf.

Wagner, Todd H., Chen, Shuo, and Paul G. Barnett. 2003. "Using Average Cost Methods to Estimate Encounter-Level Costs for Medical-Surgical Stays in the VA." *Medical Care Research and Review*. 60(3): 15S-36S.

Weiss, A. 1995. "Human Capital vs. Signaling Explanations of Wages." Journal of Economic Perspectives. 9(4): 133-154.

Werner, R.M. and D.A. Asch. 2005. "Publicly Reporting Quality Information – Reply." Journal of the American Medical Association. 293(24): 3000-3001.

Werner, R.M. and D.A. Asch. 2005. "The Unintended Consequences of Publicly Reporting Quality Information" Journal of the American Medical Association. 293(10): 1239-1244.

## Table 1: Residency Program Comparisons

| | | Program A | Program B |
|---|---|---|---|
| Affiliated Medical School Rankings (out of 126 schools): | Medical College Admissions Test (MCAT) Ranking | Top 5 | Top 50 |
| | NIH Funding Ranking | Top 5 | Top 80 |
| Affiliated Hospital | US News Honor Roll (Overall) | Top 10 | Not Listed |
| Resident Characteristics | % with MD from Top 10 Medical School (US News rankings) | 30% | 3% |
| | % with MD from Top 25 Medical School (US News rankings) | 50% | 9% |
| | % with MD from Top 10 Medical School (NIH Funding rankings) | 25% | 2% |
| | % with MD from Top 25 Medical School (NIH Funding rankings) | 40% | 8% |
| | % Foreign Medical School | 10% | 20% |
| Board Certification: Residency Program Pass Rate | American Board of Internal Medicine | 99% (95th percentile) | 85% (20th percentile) |
| | American Board of Surgery | 85% (75th percentile) | 60% (20th percentile) |

Figures are approximate but representative of rankings over the past 20 years. Sources: US News & World Report rankings, various years; American Board of Internal Medicine; American Board of Surgery; AMA Masterfile, 1993-2005

## Table 2: Summary Statistics

| | | Assigned to Program A (Odd SSN) | Assigned to Program B (Even SSN) | p-value |
|---|---|---|---|---|
| Demographics | age | 63.0 | 62.8 | 0.35 |
| | 18-34 | 0.019 | 0.022 | 0.15 |
| | 35-44 | 0.074 | 0.075 | 0.80 |
| | 45-54 | 0.186 | 0.186 | 0.94 |
| | 55-64 | 0.229 | 0.229 | 0.92 |
| | 65-69 | 0.134 | 0.131 | 0.50 |
| | 70-74 | 0.149 | 0.146 | 0.57 |
| | 75-84 | 0.179 | 0.184 | 0.39 |
| | 84+ | 0.030 | 0.027 | 0.24 |
| | male | 0.976 | 0.978 | 0.19 |
| | white | 0.466 | 0.472 | 0.42 |
| | married | 0.443 | 0.446 | 0.65 |
| | divorced | 0.271 | 0.269 | 0.80 |
| Comorbidities | Charlson index = 0 | 0.294 | 0.290 | 0.52 |
| | Charlson index = 1 | 0.274 | 0.278 | 0.37 |
| | Charlson index = 2 | 0.433 | 0.432 | 0.91 |
| Admission Time | Midnight-6am | 0.096 | 0.098 | 0.56 |
| | 6am-12 noon | 0.237 | 0.233 | 0.29 |
| | 12 noon-6pm | 0.420 | 0.425 | 0.28 |
| | 6pm - Midnight | 0.247 | 0.245 | 0.59 |
| Day of the week | weekend | 0.163 | 0.162 | 0.72 |
| ZIP Code Characteristics | median HH Income | 33714 | 33945 | 0.24 |
| | fraction HS dropout | 0.249 | 0.247 | 0.18 |
| | fraction HS only | 0.317 | 0.318 | 0.34 |
| | fraction Some College | 0.271 | 0.272 | 0.024* |
| | fraction white | 0.628 | 0.633 | 0.48 |
| | fraction black | 0.331 | 0.327 | 0.52 |
| | fraction aged 19-34 | 0.214 | 0.213 | 0.21 |
| | fraction aged 35-64 | 0.368 | 0.369 | 0.38 |
| | fraction aged 65+ | 0.141 | 0.141 | 0.22 |
| | population per 1000 sq meters | 1.102 | 1.072 | 0.09 |
| | Observations (discharges) | 35932 | 36434 | |

p-values calculated using standard errors clustered by patient. * significant at 5%;

## Table 3: Treatment Differences

| Dependent Variable: | log(length of stay) | | | log(accounting cost) | | | log(estimated expenditure) | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Assigned to | 0.108 | 0.114 | 0.113 | 0.113 | 0.123 | 0.125 | 0.100 | 0.102 | 0.104 |
| Program B | [0.0086]** | [0.0075]** | [0.0072]** | [0.0136]** | [0.0116]** | [0.0114]** | [0.0120]** | [0.0104]** | [0.0099]** |
| Diagnosis Fixed Effects | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes |
| Full Controls | No | No | Yes | No | No | Yes | No | No | Yes |
| Observations | 72366 | | | 34098 | | | 42518 | | |
| Mean of Dep. Var. | 1.43 | | | 8.63 | | | 8.71 | | |

Models estimated using OLS. Robust standard errors in brackets, clustered by patient. Full controls include variables listed in Table 1, as well as month, year, and day-of-the-week indicators. Cost measures are in 2006 dollars. ** significant at 1%

## Table 4A: Differences in VA Hospital Readmissions

| Dependent Variable: | 30-day Readmission | | | 1-year Readmission | | | 1-year Readmission Same Major Diagnosis | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Assigned to | -0.0019 | -0.0019 | -0.0021 | 0.0057 | 0.0057 | 0.0055 | 0.0032 | 0.0032 | 0.0033 |
| Lower Ranking Program | [0.0032] | [0.0031] | [0.0030] | [0.0058] | [0.0053] | [0.0051] | [0.0045] | [0.0039] | [0.0039] |
| Diagnosis Fixed Effects | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes |
| Full Controls | No | No | Yes | No | No | Yes | No | No | Yes |
| Observations | 71954 | | | 66938 | | | 66998 | | |
| Mean of Dep. Var. | 0.132 | | | 0.429 | | | 0.204 | | |

## Table 4B: Differences in Mortality

| Dependent Variable: | 30-day Mortality | | | 1-year Mortality | | | 5-year Mortality | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Assigned to | -0.0006 | -0.0006 | -0.0007 | -0.0067 | -0.0061 | -0.0072 | -0.0016 | 0.0001 | -0.0028 |
| Lower Ranking Program | [0.0020] | [0.0019] | [0.0019] | [0.0051] | [0.0045] | [0.0044] | [0.0085] | [0.0072] | [0.0068] |
| Diagnosis Fixed Effects | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes |
| Full Controls | No | No | Yes | No | No | Yes | No | No | Yes |
| Observations | 71954 | | | 66938 | | | 47337 | | |
| Mean of Dep. Var. | 0.0642 | | | 0.242 | | | 0.507 | | |

Models estimated using OLS on a sample that includes patients seen 30 days, 1 year, or 4 years from the end of the sample period. Robust standard errors in brackets, clustered by patient. * significant at 5%; ** significant at 1%.

**Table 5: Results Across Diagnoses**

| Top 10 Most Common Diagnoses | Dependent Variable | Coeff. On Assignment to Program B | S.E. | Mean of Dep. Var. | Program B Fraction | p-value: fraction=0.5 | Obs. |
|---|---|---|---|---|---|---|---|
| Heart Failure | log(length of stay) | 0.252 | [0.0272]** | 1.53 | 0.520 | 0.018 | 3598 |
| | 1-year mortality | 0.005 | [0.0210] | 0.349 | | | 3249 |
| Chronic Ischemic Heart Disease | log(length of stay) | 0.083 | [0.0299]** | 0.85 | 0.514 | 0.15 | 2662 |
| | 1-year mortality | -0.013 | [0.0125] | 0.0794 | | | 2368 |
| Acute Myocardial Infarction | log(length of stay) | 0.089 | [0.0372]* | 1.61 | 0.505 | 0.62 | 2187 |
| | 1-year mortality | -0.030 | [0.0201] | 0.248 | | | 2071 |
| Respiratory & Chest Symptoms | log(length of stay) | 0.175 | [0.0302]** | 0.77 | 0.518 | 0.092 | 2142 |
| | 1-year mortality | -0.004 | [0.0133] | 0.0914 | | | 1828 |
| Chronic Obstructive Pulmonary Disease | log(length of stay) | 0.191 | [0.0343]** | 1.36 | 0.457 | <0.001 | 2137 |
| | 1-year mortality | 0.001 | [0.0256] | 0.294 | | | 1965 |
| Diabetes | log(length of stay) | 0.131 | [0.0456]** | 1.61 | 0.544 | <0.001 | 2097 |
| | 1-year mortality | -0.025 | [0.0198] | 0.184 | | | 1920 |
| Cardiac dysrhythmias | log(length of stay) | 0.145 | [0.0392]** | 1.41 | 0.494 | 0.56 | 2034 |
| | 1-year mortality | -0.039 | [0.0205] | 0.213 | | | 1899 |
| GI Bleed | log(length of stay) | 0.163 | [0.0370]** | 1.40 | 0.493 | 0.53 | 1974 |
| | 1-year mortality | -0.015 | [0.0221] | 0.218 | | | 1856 |
| Pneumonia | log(length of stay) | 0.210 | [0.0364]** | 1.50 | 0.516 | 0.15 | 1944 |
| | 1-year mortality | 0.015 | [0.0232] | 0.307 | | | 1749 |
| Other acute and subacute forms of ischemic heart disease | log(length of stay) | 0.129 | [0.0372]** | 1.33 | 0.512 | 0.32 | 1843 |
| | 1-year mortality | -0.027 | [0.0151] | 0.0895 | | | 1821 |
| Pr(Mortality\|Diagnosis) Bottom Quartile | log(length of stay) | 0.023 | [0.0167] | 1.13 | 0.508 | 0.16 | 8767 |
| | 1-year mortality | -0.004 | [0.0047] | 0.0412 | | | 8250 |
| Pr(Mortality\|Diagnosis) 2nd Quartile | log(length of stay) | 0.112 | [0.0131]** | 1.18 | 0.510 | 0.012 | 17153 |
| | 1-year mortality | -0.008 | [0.0056] | 0.101 | | | 15765 |
| Pr(Mortality\|Diagnosis) 3rd Quartile | log(length of stay) | 0.119 | [0.0116]** | 1.48 | 0.493 | 0.030 | 26420 |
| | 1-year mortality | -0.009 | [0.0068] | 0.230 | | | 24424 |
| Pr(Mortality\|Diagnosis) Top Quartile | log(length of stay) | 0.142 | [0.0141]** | 1.72 | 0.510 | 0.0035 | 20026 |
| | 1-year mortality | -0.005 | [0.0090] | 0.466 | | | 18499 |

Top 10 most frequent diagnoses based on 3-digit ICD-9 diagnosis codes, with the exception GI bleed & COPD defined by a group of diagnosis codes. Models estimated using OLS. All models include full controls and diagnostic fixed effects. Robust standard errors in brackets, clustered b patient. *significant at 5%; ** significant at 1%.

## Table 6: Differences By Types of Care

| Dependent Variable: | Number of Procedures | Number of Surgeries | Accounting Cost Segments: Nursing | Surgery | Radiology | Lab | Pharmacy | All Other | Outpatient Referral |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Assigned to | 0.250 | -0.002 | 292 | -123 | 40 | 53 | 112 | 253 | -0.009 |
| Program B | [0.0143]** | [0.0036] | [88.2776]** | [30.5502]** | [12.1013]** | [8.8733]** | [48.6039]* | [46.0791]** | [0.0039]* |
| Observations | 72366 | 72366 | 34098 | 34098 | 34098 | 34098 | 34098 | 34098 | 72366 |
| Mean of Dep. Var. | 1.68 | 0.290 | 4145 | 1354 | 483 | 415 | 982 | 2431 | 0.793 |

Models estimated using OLS. All models include full controls and diagnostic fixed effects. Robust standard errors in brackets, clustered by patient. Cost measures are in 2006 dollars. *significant at 5%; ** significant at 1%.

## Table 7: Use of Diagnostic Tests and Non-Surgical Procedures

| Comparison: | Procedure Rate | | | # \| any | | | Days to Procedure \| ordering | | | Days to Procedure | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Program A | Program B | | Program A | Program B | | Program A | Program B | | Hazard Ratio (Program B: Program A) | S.E. | |
| | (1) | (2) | | (3) | (4) | | (5) | (6) | | (7) | (8) | |
| **A. All Cases** | | | | | | | | | | | | |
| any diagnostic | 68.4% | 73.1% | ** | 2.99 | 3.25 | ** | 1.41 | 1.55 | ** | 0.993 | 0.0069 | |
| xray | 22.4% | 25.1% | ** | 1.77 | 1.77 | | 3.04 | 3.17 | | 0.948 | 0.0075 | ** |
| chest xray | 6.3% | 7.5% | ** | 1.11 | 1.13 | * | 4.39 | 4.69 | * | 0.930 | 0.0077 | ** |
| endoscopy | 5.2% | 5.7% | ** | 1.26 | 1.30 | ** | 4.90 | 4.89 | | 0.921 | 0.0078 | ** |
| angiography | 8.1% | 8.3% | | 2.70 | 2.67 | | 3.16 | 3.53 | ** | 0.915 | 0.0077 | ** |
| cardiac stress test | 6.4% | 7.8% | ** | 1.02 | 1.02 | | 3.96 | 4.39 | ** | 0.925 | 0.0078 | ** |
| other cardiac test (incl. echo.) | 12.7% | 15.0% | ** | 1.12 | 1.11 | | 1.39 | 2.21 | ** | 0.933 | 0.0079 | ** |
| Observations | 35932 | 36434 | | | | | | | | 72366 | | |
| **B. Heart Failure** | | | | | | | | | | | | |
| any diagnostic | 78.6% | 82.7% | | 2.92 | 3.33 | * | 1.10 | 1.34 | ** | 0.937 | 0.025 | * |
| angiography | 5.6% | 6.3% | | 2.80 | 2.75 | | 4.81 | 7.26 | ** | 0.747 | 0.026 | ** |
| cardiac stress test | 11.4% | 13.6% | * | 1.03 | 1.03 | | 3.42 | 4.52 | ** | 0.771 | 0.026 | ** |
| other cardiac test (incl. echo.) | 29.7% | 33.2% | * | 1.09 | 1.15 | | 0.93 | 1.62 | ** | 0.821 | 0.027 | ** |
| Observations | 1728 | 1870 | | | | | | | | 3598 | | |
| **C. Acute Myocardial Infarction** | | | | | | | | | | | | |
| any diagnostic | 90.7% | 93.2% | * | 3.88 | 4.18 | ** | 1.26 | 1.36 | | 0.951 | 0.031 | |
| angiography | 46.6% | 46.3% | | 3.01 | 3.00 | | 3.04 | 3.36 | | 0.911 | 0.037 | * |
| cardiac stress test | 20.6% | 29.6% | ** | 1.03 | 1.03 | | 5.43 | 5.33 | | 1.010 | 0.042 | |
| other cardiac test (incl. echo.) | 33.2% | 38.0% | ** | 1.15 | 1.13 | | 2.01 | 3.02 | ** | 0.904 | 0.037 | * |
| Observations | 1082 | 1105 | | | | | | | | 2187 | | |
| **D. Chronic Obstructive Pulmonary Disease** | | | | | | | | | | | | |
| any diagnostic | 84.3% | 87.1% | | 3.26 | 3.30 | | 0.59 | 0.94 | ** | 0.909 | 0.028 | ** |
| xray | 16.0% | 18.1% | | 1.52 | 1.54 | | 2.93 | 3.58 | | 0.825 | 0.033 | ** |
| chest xray | 9.9% | 11.6% | | 1.09 | 1.07 | | 2.91 | 3.66 | | 0.838 | 0.034 | ** |
| Observations | 1160 | 977 | | | | | | | | 2137 | | |
| **E. GI Bleed** | | | | | | | | | | | | |
| any diagnostic | 75.0% | 79.4% | * | 2.68 | 2.98 | * | 0.74 | 0.94 | ** | 0.951 | 0.033 | |
| endoscopy | 59.0% | 62.8% | | 1.29 | 1.35 | * | 2.19 | 2.28 | | 0.848 | 0.034 | ** |
| Observations | 1001 | 973 | | | | | | | | 1974 | | |

Columns (1) and (2) report the fraction of patients who received the procedure at least once; Columns (3) and (4) report the number of procedures conditional on having at least one; Columns (5) and (6) report the mean number of days to the first time the procedure is conducted conditional on having the procedure; Column (7) reports hazard ratios of the duration to the first time a procedure is conducted: results are from Cox proportional hazard models with full controls. Standard errors are clustered at the patient level. * significant at 5%, ** significant at 1%

## Table 8: Specification & Robustness Checks

| | Dependent Variable | Coeff. On Assignment to Program B | S.E. | Mean of Dep. Var. | Obs. |
|---|---|---|---|---|---|
| Sample: nervous system patients | log(length of stay) | 0.047 | 0.048 | 1.34 | 1353 |
| | 30-day readmission | -0.011 | 0.022 | 0.191 | 1345 |
| | 1-year mortality | -0.040 | 0.021 | 0.153 | 1284 |
| Sample: outside main facility | log(length of stay) | -0.012 | 0.014 | 1.89 | 70775 |
| | 1-year mortality | 0.0050 | 0.004 | 0.141 | 63299 |
| Sample: first episode | log(length of stay) | 0.096 | 0.0097** | 1.40 | 29391 |
| | 30-day readmission | -0.010 | 0.0033** | 0.091 | 29278 |
| | 1-year mortality | -0.0037 | 0.004 | 0.173 | 27581 |
| | 5-year mortality | -0.0040 | 0.006 | 0.391 | 20882 |
| White veterans | log(length of stay) | 0.0759 | 0.012** | 1.48 | 33923 |
| | 1-year mortality | -0.0060 | 0.0066 | 0.239 | 33923 |
| Non-white veteran (or missing race) | log(length of stay) | 0.1380 | 0.011** | 1.39 | 38443 |
| | 1-year mortality | -0.0048 | 0.0070 | 0.245 | 33015 |
| Readmission Outcomes | 30-day readmission: same major diagnosis | -0.0020 | 0.0021 | 0.071 | 71954 |
| | 30-day readmission costs | 20.3 | 89.4 | 1653 | 42106 |
| | 1-year readmission costs | 243 | 155 | 4868 | 37090 |
| Mortality Outcomes | 10-hour mortality | -0.00042 | 0.0004 | 0.0025 | 72366 |
| | died in the hospital | 0.0020 | 0.0014 | 0.040 | 72366 |
| Transfers | transfer to another hospital | -0.0028 | 0.0016 | 0.040 | 72366 |

All models include full controls, including 3-digit diagnosis indicators. Robust standard errors in brackets, clustered by patient. * significant at 5%; ** significant at 1%.

**Table A1: Selected Covariates**

| Dependent Variable: | (1) log(length of stay) | (2) log(accounting cost) | (3) log(estimated cost) | (4) 30-day Readmission | (5) 1-year Readmission | (6) 30-day mortality | (7) 1-year mortality |
|---|---|---|---|---|---|---|---|
| Assigned to Program B | 0.1125 | 0.1251 | 0.1039 | -0.0021 | 0.0055 | -0.00073 | -0.0072 |
| | [0.0072]** | [0.0114]** | [0.0099]** | [0.0030] | [0.0051] | [0.0019] | [0.0044] |
| Midnight-6am | 0.0474 | 0.2142 | 0.1847 | -0.0175 | -0.029 | -0.0228 | -0.0401 |
| | [0.0133]** | [0.0205]** | [0.0177]** | [0.0052]** | [0.0077]** | [0.0037]** | [0.0062]** |
| 6am-12 noon | 0.1658 | 0.0808 | 0.1065 | -0.0091 | -0.0112 | -0.0098 | 0.0038 |
| | [0.0121]** | [0.0177]** | [0.0153]** | [0.0048] | [0.0071] | [0.0034]** | [0.0058] |
| 12 noon-6pm | 0.241 | 0.1297 | 0.1738 | -0.0096 | -0.0038 | -0.0046 | 0.0127 |
| | [0.0123]** | [0.0180]** | [0.0156]** | [0.0049] | [0.0074] | [0.0036] | [0.0060]* |
| Wednesday (vs. Saturday) | 0.0327 | -0.0454 | -0.0082 | -0.0018 | -0.0078 | -0.0065 | -0.0017 |
| | [0.0134]* | [0.0226]* | [0.0194] | [0.0054] | [0.0080] | [0.0038] | [0.0062] |
| Married | -0.0893 | -0.0763 | -0.07 | 0.0034 | 0.0058 | -0.0067 | -0.0264 |
| | [0.0091]** | [0.0143]** | [0.0125]** | [0.0038] | [0.0063] | [0.0024]** | [0.0056]** |
| Male | 0.061 | -0.0275 | 0.0864 | 0.006 | 0.0205 | 0.0111 | 0.0451 |
| | [0.0225]** | [0.0315] | [0.0296]** | [0.0087] | [0.0163] | [0.0042]** | [0.0129]** |
| White | 0.0158 | 0.0308 | 0.0115 | -0.0062 | -0.0004 | 0.0033 | 0.0065 |
| | [0.0112] | [0.0199] | [0.0157] | [0.0046] | [0.0076] | [0.0031] | [0.0069] |
| Charlson Index = 1 | 0.0884 | 0.0695 | 0.0974 | 0.0201 | 0.066 | 0.0032 | 0.0351 |
| | [0.0091]** | [0.0145]** | [0.0129]** | [0.0034]** | [0.0058]** | [0.0019] | [0.0040]** |
| Charlson Index = 2 | 0.202 | 0.2054 | 0.2248 | 0.0555 | 0.1422 | 0.0352 | 0.1584 |
| | [0.0099]** | [0.0158]** | [0.0140]** | [0.0039]** | [0.0063]** | [0.0025]** | [0.0053]** |
| Age: 35-44 | 0.181 | 0.1336 | 0.092 | 0.0115 | 0.0391 | 0.004 | 0.0044 |
| | [0.0295]** | [0.0659]* | [0.0500] | [0.0117] | [0.0212] | [0.0038] | [0.0137] |
| 45-54 | 0.2452 | 0.1913 | 0.1134 | 0.0101 | 0.0653 | 0.0104 | 0.0276 |
| | [0.0284]** | [0.0616]** | [0.0466]* | [0.0110] | [0.0205]** | [0.0037]** | [0.0135]* |
| 55-64 | 0.3328 | 0.2839 | 0.1319 | 0.0106 | 0.0666 | 0.0216 | 0.0621 |
| | [0.0284]** | [0.0617]** | [0.0468]** | [0.0110] | [0.0205]** | [0.0038]** | [0.0138]** |
| 65-69 | 0.3598 | 0.2533 | 0.0969 | 0.0061 | 0.0773 | 0.0303 | 0.0998 |
| | [0.0292]** | [0.0634]** | [0.0483]* | [0.0113] | [0.0208]** | [0.0043]** | [0.0144]** |
| 70-74 | 0.372 | 0.3103 | 0.1074 | 0.0111 | 0.0819 | 0.0409 | 0.1283 |
| | [0.0292]** | [0.0629]** | [0.0480]* | [0.0114] | [0.0209]** | [0.0043]** | [0.0145]** |
| 75-84 | 0.3894 | 0.2958 | 0.0775 | 0.0281 | 0.0823 | 0.0573 | 0.18 |
| | [0.0290]** | [0.0622]** | [0.0474] | [0.0114]* | [0.0209]** | [0.0043]** | [0.0145]** |
| 84+ | 0.3873 | 0.2803 | 0.0338 | 0.0164 | 0.0562 | 0.0973 | 0.3124 |
| | [0.0344]** | [0.0673]** | [0.0533] | [0.0136] | [0.0243]* | [0.0085]** | [0.0200]** |
| Constant | 1.3466 | 8.3545 | 8.6239 | 0.0388 | 0.043 | 0.0943 | 0.1759 |
| | [0.1792]** | [0.2980]** | [0.2563]** | [0.0730] | [0.1199] | [0.0484] | [0.1107] |
| Observations | 72366 | 34098 | 42518 | 71954 | 66938 | 71954 | 66938 |
| R-squared | 0.22 | 0.25 | 0.26 | 0.03 | 0.07 | 0.11 | 0.22 |
| Mean of Dep. Var. | 1.43 | 8.63 | 8.71 | 0.1315 | 0.4287 | 0.0642 | 0.2418 |

Models also included year, month, day-of-week, and divorced indicators, as well as ZIP code characteristics. Robust standard errors in brackets; * significant at 5%; ** significant at 1%

## Table A2: Additional Checks

| | Dependent Variable | Coeff. On Assignment to Program B | S.E. | Mean of Dep. Var. | Obs. |
|---|---|---|---|---|---|
| Model: Probit | 30-day readmission | -0.002 | 0.0030 | 0.133 | 71373 |
| (marginal effects) | 1-year mortality | -0.008 | 0.0048 | 0.244 | 66230 |
| Model: OLS w/ Date Fixed Effects | log(length of stay) | 0.109 | 0.007** | 1.43 | 72366 |
| | 30-day readmission | -0.003 | 0.003 | 0.131 | 71954 |
| | 1-year mortality | -0.007 | 0.004 | 0.242 | 66938 |
| Sample: Drop transferred patients. | log(length of stay) | 0.114 | 0.007** | 1.42 | 69451 |
| | 30-day readmission | -0.003 | 0.003 | 0.129 | 69047 |
| | 1-year mortality | -0.007 | 0.004 | 0.241 | 64177 |

All models include full controls, including 3-digit diagnosis indicators. Robust standard errors in brackets, clustered by patient. * significant at 5%; ** significant at 1%.

## Table A3: Effects of Experience: June vs. July

| Dependent Variable: | log(length of stay) | 30-day readmission | 1-year mortality |
|---|---|---|---|
| | (1) | (2) | (3) |
| Assigned to Program B | 0.069 | -0.0091 | 0.0025 |
| | [0.0221]** | [0.0091] | [0.0110] |
| July | -0.0008 | -0.0081 | -0.0055 |
| | [0.0213] | [0.0086] | [0.0101] |
| Assigned to Program B * July | 0.049 | 0.017 | -0.0010 |
| | [0.0302] | [0.0122] | [0.0143] |
| Observations | 12256 | 12256 | 11286 |
| Mean of Dep. Var. | 1.39 | 0.134 | 0.244 |

Sample limited to patients admitted in June or July. Models estimated using OLS with full controls. Robust standard errors in brackets, clustered by patient. * significant at 5%; ** significant at 1%.

Figure 1A: Log(Length of Stay) vs. Last Digit of SSN

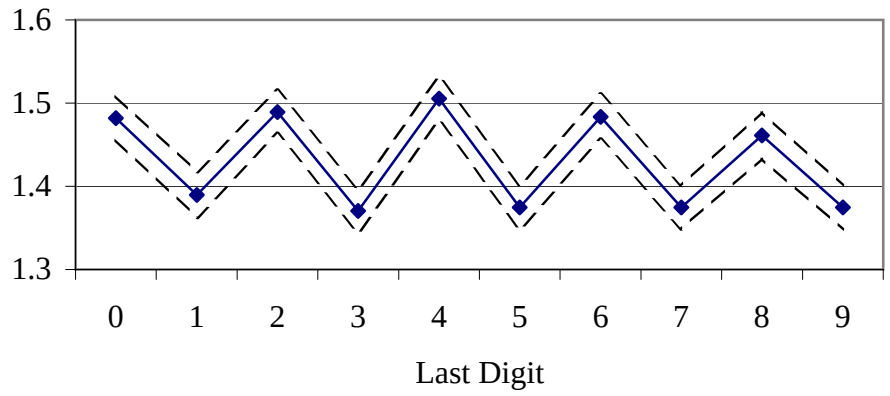


Figure 1B: Log(Accounting Cost) vs. Last Digit of SSN

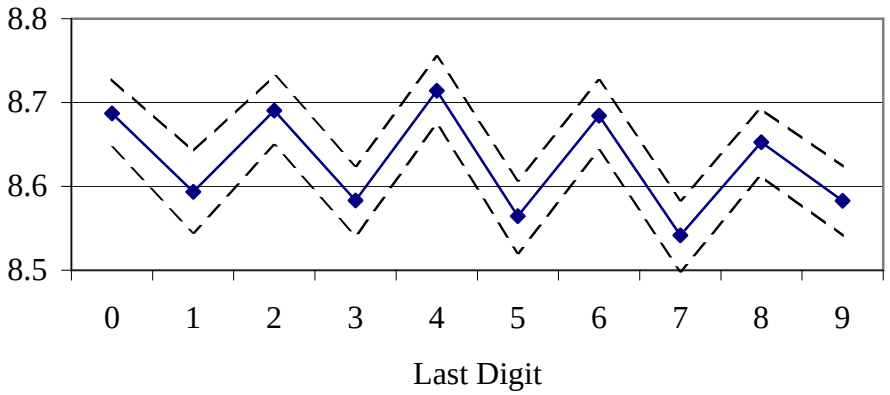


Figure 1C: Log(Est. Expenditure) vs. Last Digit of SSN

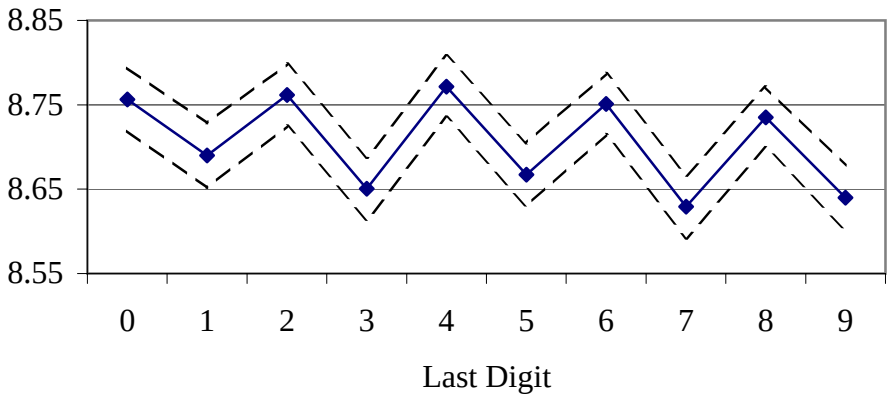

Figure 2A: 30-Day Readmission vs. Last Digit of SSN

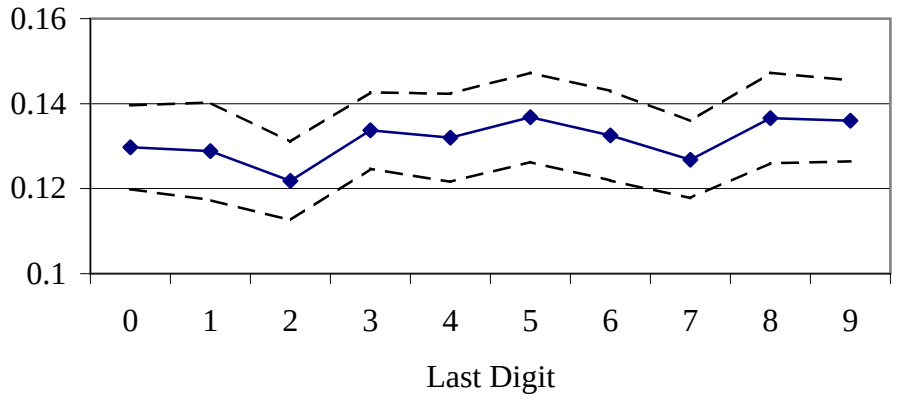


Figure 2B: 1-year Mortality vs. Last Digit of SSN

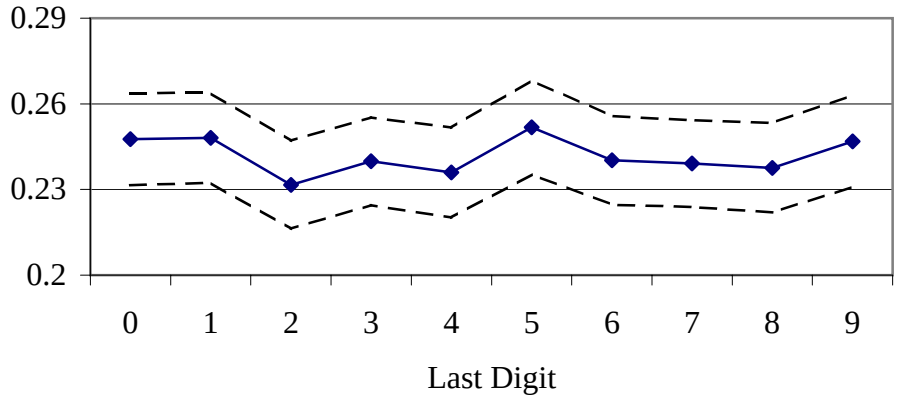


Figure 2C: 5 Year Mortality vs. Last Digit of SSN